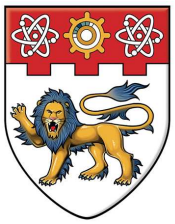
Grouping Features in Big Dimensionality



NANYANG
TECHNOLOGICAL
UNIVERSITY

A thesis
submitted to the Nanyang Technological University
in fulfilment of the requirement
for the degree of
Doctor of Philosophy of

ZHAI Yiteng

August 2015

Abstract

To date, the world continues to generate quintillion bytes of data daily, leading to the pressing needs for new efforts in dealing with the grand challenges brought by Big Data. When talking about big data, there is a growing consensus among the computational intelligence communities that data volume presents an immediate challenge pertaining to the scalability issue. Note that, when addressing volume in Big Data analytics, researchers have largely taken a one-sided study which refers to the "Big Instance Size" factor of the data. The flip side of volume which is the "Big Dimensionality" factor of big data, on the other hand, has received much less attention. A motivating example is related with cell phone manufacturing industry. It is worth noting that presently one can easily enjoy up to an extremely high resolution of 41-megapixels on the pictures taken, which is 400 times more than the 0.11-megapixels almost a decade ago. As a pixel based feature representation, this will explicitly translate to 41 million features.

Taking this cue, in this dissertation, the first work represents an attempt to fill in this gap and places special focus on this relatively under-explored topic of big dimensionality, wherein the explosion of features brings about new challenges to computational intelligence. An analysis of three popular data repositories has uncovered an exponential increase in the dimensionality of many datasets that have been produced since early 2000s, there is much evidence reinforcing our contention that the upward trend of Big Dimensionality will only continue to follow, as influenced by the *rapid advancements in computing and information technologies* and the *arising myriads of feature descriptors*. Moreover, the *blessings of Big Dimensionality* is also discussed based on feature correlation, which serves as a cue to the success of handling such challenge.

Based on the observation of the growing trend of big dimensionality on modern databases, existing approaches that require the calculations of pairwise feature corre-

lations in their algorithmic designs have scored miserably, since computing the full correlation/covariance matrix (i.e., square of dimensionality in size, where million features would translate to trillion correlation computations) can become computationally very impractical. This poses a notable challenge that has received little attention in the field of machine learning and data mining research. Thus, an efficient feature grouping and selection method has been proposed to fill in this gap, which is considered as the second work presented in this thesis. Specifically, the interesting findings on several established databases with big dimensionality have indicated that an extremely small portion of the feature pairs could contribute significantly to the underlying interactions and there exists feature groups that are highly correlated, which is termed as "sparse correlation" in this thesis. Inspired by the intriguing observations, a novel learning approach, namely, Group Discovery Machine (GDM) is hence introduced that exploits the presence of sparse correlations for the efficient identifications of informative and correlated feature groups from big dimensional data that translates to a reduction in complexity from $O(m^2n)$ to $O(m \log m + \mathcal{K}_a mn)$, where $\mathcal{K}_a \ll \min(m, n)$ generally holds. In particular, the proposed approach considers an explicit incorporation of linear, nonlinear or specific correlation measures as constraints in the learning model. An efficient embedded feature selection strategy, designed to filter out the large number of non-contributing correlations that could otherwise confuse the classifier while identifying the correlated and informative feature groups, forms one of the highlights of this approach. Extensive empirical studies on both synthetic and several real-world datasets comprising up to 30 million dimensions are subsequently conducted to assess and showcase the efficacy of the proposed framework.

In addition, to better illustrate the properties of the proposed framework, the sensitivity analysis of the key parameters in GDM are examined in this thesis to demonstrate the robustness and stability. Besides, the proposed framework on different machine learning settings is discussed, such as one-class learning, where notable speedup can be observed when solving one-class problems on big dimensional data. Further, to identify robust informative features with minimal sampling bias, the embedding of the V-fold cross validation in the learning model is hence considered, so as to seek for features that exhibit stable or consistent performance accuracy on multiple data folds. Last but not least, to

better illustrate the usefulness of the informative feature groups, the potential benefits of affiliated features are presented using various real-world datasets.

Acknowledgments

First of all, I would like to express the deepest gratitude to my main academic supervisor, **Prof. Yew-Soon Ong**. His sincere encouragement during my indecisive moments and inspirational insights into the many aspects of my research direction have faithfully guided me throughout my research work. Without his guidance, the achievements of my research work will never be possible. Moreover, I would also like to show my great appreciation to my co-supervisor, **Prof. Ivor Wai-Hung Tsang**, whose expertise, patience, and understanding, added iridescent colours to my graduate experience. I appreciate his vast knowledge and skill in the machine learning area, and I sincerely hope him a good life journey in Australia.

Furthermore, I would like to thank all my fellow and colleagues for the stimulating discussions, for the sleepless nights we were working together for deadlines, and for all the fun we have had in the past years, particularly Dr. Feng Liang, Dr. Jiang Siwei, Dr. Abhishek Gupta, Dr. Ramon Sagarna, Dr. Mao Qi, Dr. Chen Xianshun, Dr. Giduthuri Sateesh, Dr. Mostafa Mostafa Hashim, Dr. Iti Chaturvedi, Mr. Guo Ruiliang, Mr. Hou Yaqing, Mr. Tang Jing, Mr. Da Bingshui, Mr. Ali Alizadeh Mansouri, Mr. Chen Caishun and Mr. Choon-Sing Ho. Working and living with them inspires my research and makes the Ph.D. life lively and interesting.

My sincere thanks also goes to the other members of the research committee, and the staffs of not only the Centre for Computational Intelligence (C^2i) but also the MultiplAtform Game Innovation Centre (MAGIC Labyrinth) for their technical supports. Special thanks go to Mr. Kesavan Asaithambi, Mr. Zay Yar Aung and Mrs. Linda Ang for their great help in time during my research life.

Last but not least, my immense gratitude goes to my parents for the constant support, great love and continuous encouragement through my entire life. Also I show my great

thanks and love to Ms. Huang Jing for all the support she did for my study. Without these, I would not have finished my Ph.D. works easily.

Contents

List of Figures

List of Tables

Publications

- **Yiteng Zhai**, Yew-Soon Ong, Ivor W. Tsang. "Making Trillion Correlations Feasible in Feature Grouping and Selection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. *accepted*.
- **Yiteng Zhai**, Yew-Soon Ong, Ivor W. Tsang. "The Emerging 'Big Dimensionality'". *IEEE Computational Intelligence Magazine*, 9(3):14-26, August 2014.
- **Yiteng Zhai**, Mingkui Tan, Ivor W. Tsang, Yew-Soon Ong. "Discovering Support and Affiliated Features from Very High Dimensions". *in Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, (Edinburgh, Scotland, United Kingdom), pp. 1455-1462, Omnipress, 2012.

Chapter 1

Introduction

1.1 Motivation of the Research on Big Dimensional Data

Machine learning is a subfield of artificial intelligence, with the focus placed on healthy development of robust algorithms and techniques that allow computers to learn. It has a wide spectrum of applications such as cheminformatics, search engines, medical diagnosis, stock market analysis, game playing and more recent hot research fields including bioinformatics, computer vision and natural language processing.

Over the last decade, an exponential growth in the dimensionality of the datasets has been witnessed across many domains [59–61], hence leading to a new level of scalability study of machine learning approaches in this new era of Big Data. Note that, researchers in the data analytics community have largely taken a one-sided study of data volume [62–64], which refers to the instance size of the data. The corresponding factor of "Big Dimensionality", on the other hand, has received much less attention in the context of Big Data Analytics [65]. Take bioinformatics for example, the amount of biological data requiring analysis has ballooned and many new and specialized machine learning methods have since emerged to deal with this explosion of data. Consequently, machine learning in bioinformatics has become an important research interest of both computer scientists and biologists. As a way to protect end users from visiting undesirable sites, another notable application would involve the classification between malicious web sites (involved in criminal scams) and benign sites, using the lexical and host-based features

of the associated URLs [66, 67]. Typically, the individual URLs are encoded[1] as big dimensional feature vectors[2]. As such systems involve the gathering of URLs crawled at real-time, from the World-Wide-Web, fast prediction to fulfill the requirement of low latency is necessary. Thus a subset of the original features or feature groups is desirable for accelerating the learning process while maintaining high classification accuracy. To this end, seeking an efficient method to robustly reveal the intrinsic feature structure of the dataset is of urgent need.

1.2 An Established Therapy for "Curse of Dimensionality" – Feature Selection

In the past decade, many real-world datasets are thus represented with very high dimensional features, bringing about significant challenges in the data mining field [68, 69]. Learning performance is often degraded with the inflation of dimensions, leading to the well-known notion - "Curse of Dimensionality" [70, 71], which brings great challenges in data mining and processing [71–73]. Fortunately, the research on the big dimensional data also reveal that most of the features are irrelevant to the output (i.e., noisy features). To address this issue, a well established remedy is Dimension Reduction, which is the process of reducing the number of random variables under consideration. And the most well-known subarea in this field is **Feature Selection**.

Feature selection focuses on seeking a small feature subset among the **original** dimensions that is most relevant to the task label. Although it is deemed as an old topic in pattern recognition, feature selection is known to be very effective on many machine learning tasks. For example, in high dimensional data such as text data, many features are usually non-informative or noisy (i.e., not every word or word combination could be equally predictive), resulting in serious deterioration of the generalization performance, whereas feature selection is deemed as a notable remedial tool for addressing the issue [74]. Further, a sparse classifier can often lead to simplified decision rules that offer faster yet accurate prediction on large-scale problems [75]. Last but not least, on many real

[1] In URL representation, most of the features are generated by the "bag-of-words", where '/', '?', '.', '=', '-', and '_' are delimiters.

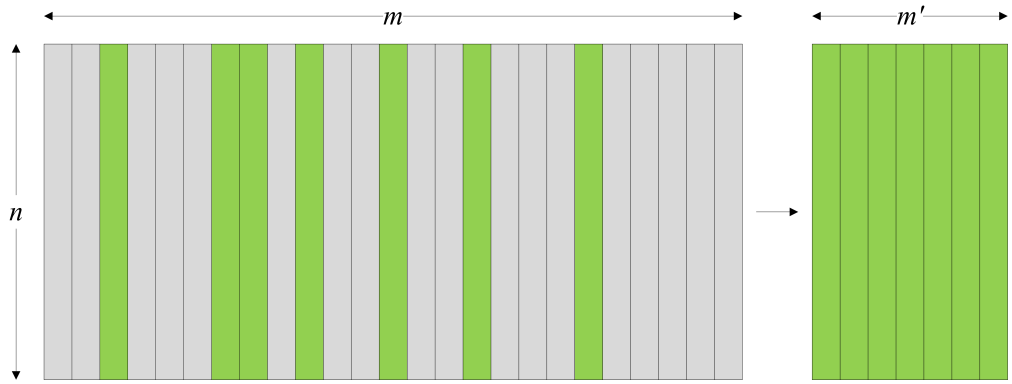
[2] Suspicious URLs proposed in [67] has approximately 3 millions in dimensionality.

world applications such as microarray data analysis, a small set of input features is typically desirable to enable better interpretations of the results. As such, to date, feature selection is recognized as one of the most important tasks in machine learning research and has remained to be studied by many researchers in the recent years.

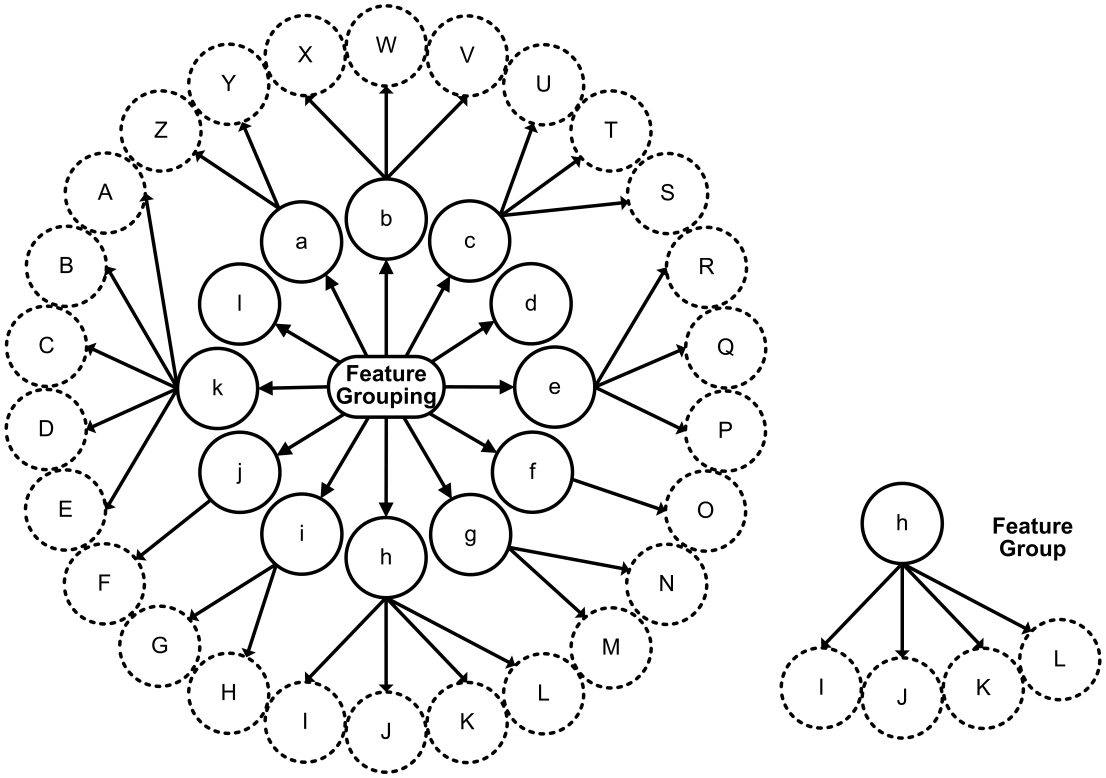
1.3 More Robust Learner – Feature Grouping

Besides the idea on purely reducing the dimension, there are increasing interests on identifying structures in the data [76, 77]. To date, it is worth noting that most of the existing feature selection methods generally assume a *good feature subset* [78] as one that possesses strong prediction ability pertaining to the output labels; meanwhile the selected features should also maintain low correlations among themselves (i.e., low pairwise feature correlation). In other words, each chosen feature is desired to be a carrier of both significant and unique information. Correlated features however are typically deemed as redundant, and such redundancy should be minimized [78–81]. Though eliminating redundant features has been widely used in practice and regarded as the guiding principle behind the development of feature selection methods, it may not always hold since these correlated features can be useful (i.e., informative) for the tasks on hand as part of the structure that represents the dataset. As also discussed in [82–84], such feature redundancy (i.e., represented by high feature correlation) may have the benefits of bringing about stable generalization performances.

Taking this cue, **Feature Grouping** is introduced as the technology that has been developed to gather correlated informative features into different functional feature groups, which can be further gathered to form the feature structure/network. Specifically, the difference between feature grouping and feature selection lies in the outputs, where a feature grouping method returns feature groups/clusters or even the overall feature structure, rather than the indice set that provided by a feature selection method, as depicted in Figure 1.1 (instead of outputting the subset of features, feature grouping methods aim to find the feature groups using some specific feature dependency measures, and the chosen feature groups can be further gathered as a structure as what Figure 1.1.b shows). Over the past few years, feature grouping has been demonstrated to be promising over feature



1.1.a: Illustration of Feature Selection



1.1.b: Illustration of Feature Grouping

Figure 1.1: a: Task of Feature Selection on dataset $\mathbf{X}$ with m number of features and correspondingly each of n attributes. Each bar denotes corresponding feature while the green colour denotes the selected features. b: Illustration on the structural relationship of Feature Grouping and Feature Group, where full circle along with the linked dotted circles form feature group as output.

Figure 1.2: Visual difference in human understanding between feature selection and feature grouping, by differentiating a gentleman and a lady using `The ORL Database of Faces`. Note that, the white pixels denote the selected features. (Left: original face images; Middle: face images with feature selection outputs; Right: face images with feature grouping outputs.)

selection, since it can help to reduce the variances in the estimation and improves the stability of feature selection and gain additional insights to understand and interpret data [85, 86].

Here, images from `The ORL Database of Faces`[3], as depicted in Figure 1.2, are deemed as a suitable example to illustrate the differences between these two technologies. Specifically, a perfect prediction by the classifier on male and female face images can be attained based on the "optimal features" identified using the feature selection method, as depicted by the white pixels (which are observed to be sparsely spread across the entire image) in the middle column of Figure 1.2. However, this brings about little insight that may assist the human user in the interpretation of the features. On the other hand, the white pixels shown in the right column of Figure 1.2, denoting some insightful feature groups, are much more informative in understanding the task. To be precise, with such feature groups, it is easy for the human user to spot the core group features in regions such as mustache and beard, even the silhouette of the face, that can be helpful in grasping a better understanding of the critical objects in the images. Recall that, existing correlation based feature selection methods usually eliminate these correlated features and treat them as redundancies.

[3]http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

1.4 Objectives

Over the last decade, there has been an exponential growth in the dimensionality of the datasets that were generated. To address this trend, there is a pressing need for new ways in coping with the unprecedented data dimensions that are scaling to levels that now renders existing computational intelligence approaches inadequate. However, to date, there remains to be a lack of comprehensive studies in the literature that analyses this emerging trends of Big Dimensionality in the era of Big Data. Taking the cue, the first objective of this thesis is set out to fill in the gap for a comprehensive survey on Big Dimensionality. Secondly, this research introduces feature grouping as a noteworthy manner to mitigate the curse of dimensionality. The identified feature groups have been shown to contain substantive characteristics of the features, providing functional interpretation of the features to the prediction task well. At the same time, feature groupings can be useful to assist users in their interpretations of the data for further analysis. Further, attempts are made to acquire correlated informative features in the process of forming the feature groups more efficiently. This is in contrast to existing feature grouping approaches which have scored miserably in computations due to the need to perform calculations of the correlation/covariance matrix in their algorithmic designs, which involves computing the squared of dimensionality.

1.5 Organization of the Thesis

In what follows, the organizations for the rest of this thesis are described.

- **Chapter 2 – A Survey on The Emerging "Big Dimensionality".** This chapter presents a thorough survey on the emerging problem about big data from the perspective of dimension. Specifically, as a growing consensus among the computational intelligence communities, volume of big data (i.e., big instance size) presents an immediate challenge pertaining to the scalability issue, however, the flip side of dimensionality factor of big data, on the other hand, has received little attention. This chapter thus represents an attempt to fill in this gap and places special focus on this relatively under-explored topic of "Big Dimensionality", wherein the explosion of features brings about new challenges to computational intelligence. Firstly,

the origins of big dimensionality is analysed. Subsequently, the evolution of feature dimensionality in the last two decades is studied using popular data repositories considered in the data analytics and computational intelligence research communities. Last but not least, the "curse and blessing of big dimensionality" are delineated and deliberated.

- **Chapter 3 – Literature Review.** The objective of this thesis, as illustrated in Chapter 1, is to propose a technique that can group correlated and informative features of big dimensionality for classification problems. Accordingly, this chapter serves the purpose of providing a rigorous background study of the problem statement. Specifically, the essential conceptions of machine learning problems shall be introduced first. The emerging technique of feature grouping shall then be discussed in detail, followed by an introduction to some state-of-the-art methods. Further, to adapt the idea of feature grouping to big dimensional data, a viable approach is that of extending existing feature selection methods that can handle big dimensionality and feature correlation. Taking this cue, state-of-the-art feature selection methods are consequently reviewed using a traditional taxonomy. In addition, to justify the aim of choosing features from the original feature space, a comparison between feature selection and feature extraction is also clarified in this chapter.

- **Chapter 4 – Group Discovery Machine.** The survey on "Big Dimensionality" (i.e., Chapter 2) has highlighted that modern databases with big dimensionality are showing a rapidly growing trend. Consequently, from the perspective of feature grouping, existing approaches that require calculations of pairwise feature correlations in their algorithmic designs have scored miserably on such databases, since computing the full correlation/covariance matrix (i.e., square of dimensionality in size) can become computationally very intensive. With this in mind, this chapter thus describes the proposed study which attempts to fill in the gap. The findings on several established real world datasets have indicated that an extremely small portion of the feature pairs contributes significantly to the underlying interactions, and there exists feature groups that are highly correlated. Inspired by these intriguing observations, a novel learning approach – Group Discovery Machine (GDM) is

introduced, which exploits the presence of sparse correlations for the efficient identifications of informative and correlated feature groups from big dimensional data that translates to a reduction in complexity from $O(m^2n)$ to $O(m \log m + \mathcal{K}_a mn)$, where $\mathcal{K}_a \ll \min(m, n)$ generally holds. In particular, the proposed approach considers an explicit incorporation of both linear and nonlinear correlation measures as constraints in the learning model. An efficient embedded feature selection strategy, designed to filter out the large number of non-contributing correlations that could otherwise confuse the classifier while identifying the correlated and informative feature groups, forms one of the highlights of this approach.

- **Chapter 5 – Further Discussions and Problem Settings of the GDM Framework.** To better illustrate the properties of the proposed GDM framework, the sensitivity analysis of the key parameters in GDM are firstly examined in this chapter to demonstrate the robustness and stability. In addition, the proposed GDM framework on *one-class learning* is discussed, where notable speedup can be observed when solving one-class problems on big dimensional data. Further, to identify robust informative features with minimal sampling bias, the embedding of the V-fold cross validation into the learning model is considered, so as to seek for features that exhibit stable or consistent performance accuracy on multiple data folds. Last but not least, to better illustrate the usefulness of the informative feature groups, the potential benefits of affiliated features are presented using various real-world datasets.

- **Chapter 6 – Concluding Remarks and Future Work.** In this chapter, the works presented in this thesis are concluded. In addition, the potential future works are also outlined from several aspects.

Chapter 2

A Survey on the Emerging "Big Dimensionality"

2.1 Introduction

As we embark on the new era of Big Data, many industrial leaders today are earnestly seeking for new ways to enhance and empower consumer experiences, increase productivity and sales, through making sense of the data that is now becoming ubiquitous. Grasping the fact that a majority of the data generated in the world have been produced within the last two years while human continue to create quintillion bytes daily [87–89], there is a real pressing need for credible research into large-scale data analytics. This has led to the rising number of researchers that devote much time and efforts in dealing with the challenges brought about by Big Data. In recent years, the core challenges of Big Data have been widely established and can be summarized under the popular 5**V**s in Table 2.1 [62–64, 90].

From a survey of the literature, there is a growing consensus among data scientists that each "**V**" brings about unique challenges to the overall task considered in Big Data Analytics. For instance, **volume** presents the immediate challenges pertaining to the scalability issue of Big Data. Also, this is what directly comes to one's mind when referring to the term "Big". However, it is worth highlighting that, researchers in the data analytics community have largely taken a one-sided study of **volume** [62–64], which refers to the "Big instance size" factor of the data; the corresponding factor of "Big Dimensionality", on the other hand, has received much less attention in the context of

Table 2.1: 5**V**s for Big Data analytics.

Volume	Refers to the massive amounts of data that have been generated across a wide range of sources. In the context of data analytics, volume can be regarded as the product of instance size and dimensionality of the data.
Velocity	Refers to the rate when considering the acquisition and update of data. The fast-moving data thus yields an imperative need for frequent decision making using reliable computational intelligence (CI) technology.
Variety	Refers to the presence of multiple data types (or different feature representations) that take roots in various sources and range from text, image, audio, video, social media, to web logs, and so on.
Veracity	Considers the diverse quality and security levels of the data. Unhealthy sparsity, missing attributes and incomplete data are some of those considered under this category of interest.
Value	Refers to the benefits that can be gained from analysis on Big Data and the value of insights that can be extracted from Big Data. Very often, industrial players and researchers regard this as the core motivation and driver behind the study of Big Data analytics.

Big Data analytics [65]. To date, some studies on *high dimension small sample size* problems have been reported, such as random projection, Naïve Bayes, and others [91–94]. Theoretical efforts on Big Data with millions of dimensions have, however, remained relatively under-explored. In contrast to previous studies, in this chapter, an attempt has been made to fill in the gap by putting focus on this under-explored topic of Big Data analytics – "Big Dimensionality", wherein the explosion of features brings about new challenges to computational intelligence. In what follows, a peek through various stages of *feature life cycle* begins this study, namely feature description, feature selection and feature evaluation, as depicted in Figure 2.1. For each of the 3 stages (i.e., A, B and C in Figure 2.1), a brief overview that focuses on the flip side of **volume** in big data is presented – by placing the spotlight on the emerging phenomenon of Big Dimensionality and the challenges ahead.

2.1.1 Stage A: When Features Are Born

There are many ways to solve problems, and as many ways to describe them. Thus, the means of generating features will be markedly different due to the different representations that are of interests. Depending on one's experience, knowledge and understanding of the domain, a **variety** of feature description methods and representations can be introduced. For example, when working with images, depending on one taking a bird's

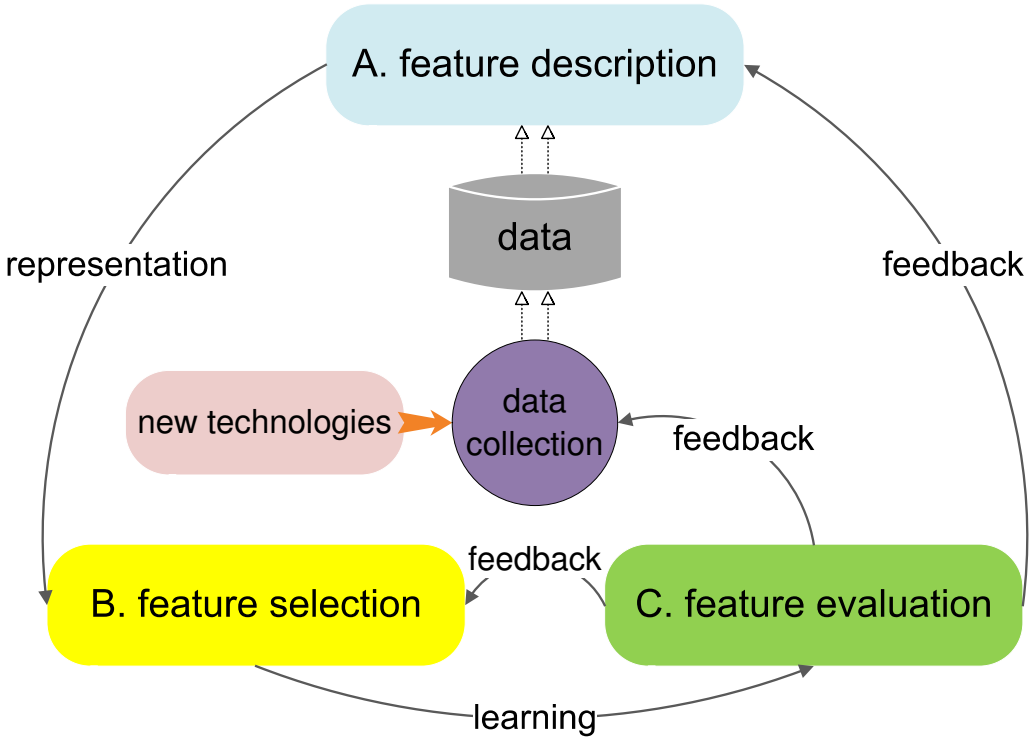


Figure 2.1: Feature Life Cycle.

eye view, a worm's eye view or an eagle's eye view, global or local features can be derived and represented with a multi-view. This makes image feature description complex in structure, while abundance in **volume** and **variety**. Even for the evening news that people are endowed with daily, researchers and engineers working in the background have to face with similar challenges of operating with the different languages in presentation. In the field of natural language processing (NLP), practitioners have to work with multiple feature types such as words, bigram/trigram templates, part-of-speech tagging templates, etc., simultaneously, in order to arrive at comprehensive representations that produce reliable predictors [43]. This myriad of feature types and their mixture are becoming a norm in many of today's real world applications [95] and together with the rapid advancements in computing and information technologies, they are major contributors of feature explosion and continue to fuel Big Dimensionality.

2.1.2 Stage B: Features Are Alike, Features Are Different

Everything has two sides, hence it is natural for some features to contribute alike, while some features differently. Since it is often the case that not all features carry equal weights on the prediction models and the application domains of interest, *feature selection* is the process of retrieving a subset of relevant features from the original feature space, for the purpose of building robust, accurate and fast learning models. Serving as the enabler for fast and cost-effective predictors, while keeping checks on the requirements in measurement and storage, feature selection is often regarded as one of the most important tasks in Big Data research. The emerging phenomenon of Big Dimensionality however calls for fresh feature selection strategies that are capable of coping with the explosion of features. Particularly, one has to deal with the explosive combinatorial effects of features or the *"curse of Big Dimensionality"*, while seeking to identify a good feature subset that is of high **value** from the original "Big" set comprising potentially irrelevant, redundant, noisy and missing features (uncertain). Due to the high importance of feature selection and many significant challenges that this stage is imbued with, it has remained and will remain to be one of the core research interests in the fields of computational intelligence and Big Data research[1].

2.1.3 Stage C: Assessing A Good Feature Subset

Assessment is particularly essential to guide the search towards high **value** features in Big Data. Every search algorithm, except for uniform random search, introduces some kind of bias into its search. Different performance metrics used in the search for evaluations exhibit unique biases [99]. It is these biases that lead the search towards particular subset of features that differs from the others. In the last decade, multivariate performance measures have been regularly introduced for the reliable *evaluations* of features, leading to highly complex criteria for assessing predictive models [100]. Further, to *verify the authenticity* of the identified features for the application domains of interest, the

[1]The *IEEE International Conference on Data Mining (ICDM) 2013* and *IEEE World Congress on Computational Intelligence (WCCI) 2014 & 2016*, organize workshops/special sessions that focus on the the issues of high dimensionality [96–98], while seeking for notable feature selection strategies that are capable of addressing the *"curse of big dimensionality"* and uncovering the potential *"blessing of big dimensionality"*.

availability of specialized human experts that are equipped with appropriate domain knowledge are essential. In this regard, a key technology that is helpful to human experts in data analytics is *visualization*. The presentation of data in different visual forms such as graphs, diagrams, charts, maps and other specialized means, can lead to easier and faster capturing of critical information as well as enhancing human understandings. Thus the technology that pushes the field forward would rely on the visualization of features, wherein a proper presentation can help in isolating the **values** of the features and subsequently figuring out the potential directions for further developments.[2] However, traditional feature verification and visualization approaches are likely to become obsolete in the face of big dimensionality.

2.1.4 Contribution of the Survey

From this survey, it is noted there has been a lack of studies that focus explicitly on analyzing the emerging trends of Big Dimensionality in the era of Big Data. The objective of this survey is specifically set out to fulfill such a role. This study begins by concentrating on the influences of *advancing technology* and the arising *myriads of feature descriptors* on the origin of big dimensionality. An analysis on the evolution of feature dimensions is then conducted based on popular data repositories used in the data analytics and computational intelligence research communities. Subsequently, a review on the state-of-the-art feature selection methods in the field of computational intelligence reveals the insufficiencies of existing approaches in keeping pace with the explosion of dimensionality. Based on the analyses, the "curse and blessing of big dimensionality" are delineated. It is worth noting that such a study would be informative to the data analytics and computational intelligence communities since it underlines the emerging trend of big dimensionality and deliberates on the curse and blessing of such developments. Further, it is hoped that the acknowledgement on big dimensionality will serve to promote the need for greater research efforts in the subject and assist in identifying new important research directions.

[2]In the past, various new feature descriptors are inspired by feedbacks obtained from the verified models and visualized results [101, 102].

2.2 The Origin of Big Dimensionality

In this section, the key factors that accounts for the origin of Big Dimensionality is focused. Specifically, a study on the origin of features is firstly presented, by focusing on how they come about, how they are represented and then reveal the core bases for the upsurge in feature dimensions over the recent years as a result of the *advancements in technology* and the *myriad of feature descriptors* that have emerged. Here, the domain of image and video learning is further showcased, since it is a popular domain of computer science as motivated by the rising popularity of the Internet and mobile devices. Subsequently, some new insights into the evolution of feature size (dimension) through analyzing three widely used data repositories of the data analytics and computational intelligence research communities are presented.

2.2.1 Advancements in Technology

Today, the advancement in computing and information technologies is happening at a rate that is far beyond our expectations. In the cell phone manufacturing industry for example, the technology of the embedded camera is progressing by many factors each year as fuelled by the innovations in diverse avenues; ranging from the flashlight, processor, sensor size, photosensitive element to the operating system and technics (*e.g.*, optical image stabilizer[3], PureView[4], etc.). The significant developments in the area of smart devices and image processing tools are empowering consumers with the capacity to generate extremely high resolution photos and video captures at anytime and anywhere, with great ease.

In the *"Cell Phone Activities 2012"* annual report [103], the portion of cell phone owners that uses their phone to take pictures was reported to have reached an astronomical rate of 82%, which is incidentally also the highest among all activities[5] made on the phone in 2012. This is a rise of 6% from 2010 on such activities. With the growing

[3] `http://www.usa.canon.com/cusa/consumer/standard_display/Lens_Advantage_IS`

[4] `http://i.nokia.com/blob/view/-/2000652/data/4/-/Download-pureview-920.pdf`

[5] These activities include picture taking, sending/receiving text messages, accessing the internet, sending/receiving email, video recording, apps download, searching for health or medical information online and online banking.

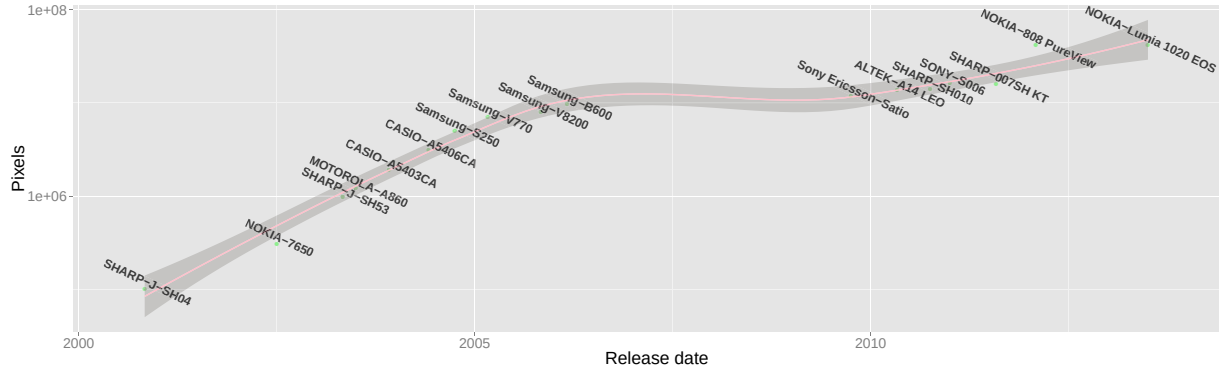


Figure 2.2: A depiction on the evolving Pixel size of images produced by a series of advancing cell phone embedded cameras, beginning from the Year 2001 to present.

popularity of online social networking services and free social media applications, including Facebook[6] and Flickr[7], photo-taking activities over the cell phone are expected to continue expanding at an accelerating rate.

When working with images, pixel is the basic cornerstone of data dimension considered by most feature descriptors (also known as feature generator or feature detector) and processing algorithms (e.g., deep learning). Figure 2.2 summarizes the growth in the feature dimensions with respect to the pixel size (y-axis) and the resolution of the embedded-camera in cell phone, from the Year 2001 (i.e., this is the year for the birth of camera phone) to present (x-axis). From the figure, there is a clear distinct indication of an exponentially increase in the feature dimensions (pixels) of camera phone generated images over the last decade. With respect to the NOKIA Lumia 1020 EOS, for instance, it is worth noting that presently people can easily enjoy up to an extremely high resolution of 41-megapixels on the pictures taken, which is 400 times more than the 0.11-megapixels image produced by the SHARP J-SH04 almost a decade ago. With much evidence that the rise in the dimensionality of data representation is set to continue, the emergence of Big Dimensionality is expected to further intensify the challenges in Big Data.

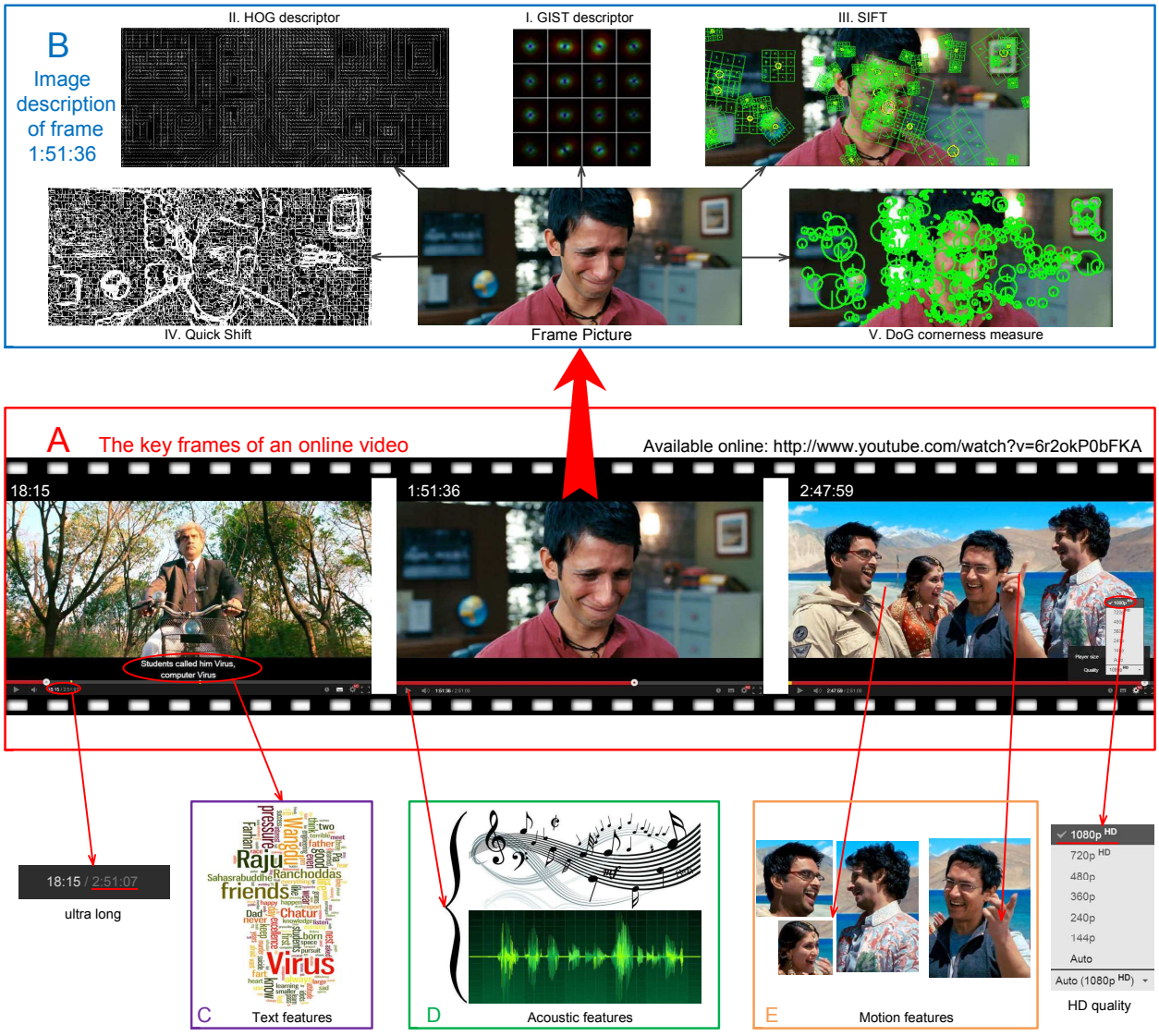


Figure 2.3: An illustration on the myriads of features descriptors for video content. A: The key frames of an online video; B: The different feature descriptors of image; C: Text features; D: Acoustic features; E: Motion features.

2.2.2 Myriad of Feature Descriptors

The advancements in computing devices and information technology have also led to new creations of features and representations, some of which can be highly detailed and sophisticated. In this Internet and mobile device dominated era, our way of life has been greatly influenced by the wide variety of media, services and applications that are now readily available online. It has been reported that multimedia content including text, image, 3D graphics, audio and video accounts for over 60% of traffic in the Internet [104].

Today, the escalation of users that enjoy spending their leisure time watching videos, browsing photos and sharing them online continue to drive the research efforts on data-centric media computing platforms that are capable of performing large-scale automatic analysis, understanding, summary, collection, organization, query and searching of multimedia content. Among the wide range of multimedia, online video has been found to account for more than half of the Internet traffic. Recently, the *YouTube Statistic* reported that 100 hours of video clips are constantly uploaded within every minute [105].

In the last decade, the pace in technological innovations of media formats and descriptors of video has risen significantly and there is much evidence that this trend will continue to rise. Video format, in particular, has evolved with increasing definition, resolution and content, advancing from 1080i, 720p, 1080p to 4K in a short period of only 5 years. In order to facilitate high accuracy learning of sophisticated video content, a myriad of feature descriptors have been introduced. In what follows, some of the core feature descriptors of online video content that contribute to the explosion of data dimensionality are discussed.

Figure 2.3A, for example, shows three key frames of an online video taken from the online social media service provider, YouTube[8]. Each of the frame can be processed separately in the form of an image, as illustrated in Figure 2.3B. In image processing, from basic pixel features, researchers have embarked on the development of complex descriptions as an important step for further analysis. GIST or simply *Spatial envelope*,

[6] `https://www.facebook.com`
[7] `https://www.flickr.com`
[8] `http://www.youtube.com/`

for instance, was introduced as a holistic descriptor that captures the core objects in the pictures [106], see Figure 2.3B-I. *Histogram of oriented gradients* (HOG), on the other hand, generates image features based on the gradient information of small cells that have been segregated in the image [107], as shown in Figure 2.3B-II. As HOG exhibits a silhouette of the original image, it is widely used for object detection of static imagery. Of equal importance is the *Scale-invariant feature transform* (SIFT) descriptor, which has been designed for image mapping, conducts a point-matching between different views of the same scenes [108]. As SIFT is invariant to translations, rotations and scalings in the image domain and robust to moderate perspective transformations and illumination variations, it is popular in the computer vision community, see Figure 2.3B-III. Subsequent extensions of SIFT include the *Speeded up robust features* (SURF) [109] and the *Gradient location and orientation histogram* (GLOH) [110] descriptors, which were designed for gains in speed and prediction accuracy, respectively. Other popular image descriptors include the *Quick shift* [111] and the *Difference of gaussians cornerness measure* [112], etc., whose feature representations are depicted in Figure 2.3B-IV and Figure 2.3B-V, respectively.

In addition to the image features, Sub-figures 2.3C, 2.3D and 2.3E depict the other typical forms of descriptors considered in online video learning [113–116], which include the motion information, audio information, (*i.e.*, acoustic feature families including *Mel-frequency cepstral coefficients*) and text information (*i.e.*, derived from the scripts and subtitles of video that were inserted by human), respectively.

With the ongoing surge in demand for enhanced user experiences and services over the Internet and mobile platforms, the pressure for highly accurate and fast processing of multimedia content involving myriads of feature descriptors can only continue to grow. As discussed, the rapid advancements in digital sensors have given birth to video images that can contain up to 4K resolution. With the myriads of feature descriptors that are available for representing video contents (*i.e.*, image, motion, acoustic and text), many millions of features or dimensions could easily transpire. Such a trend is non-isolated and can already be observed across many branches of applications. In life science research, for example, the search for a compact gene subset comprising tens of relevant biomarkers (genes) from the original thousands in microarray data is deemed as crucial to biologists

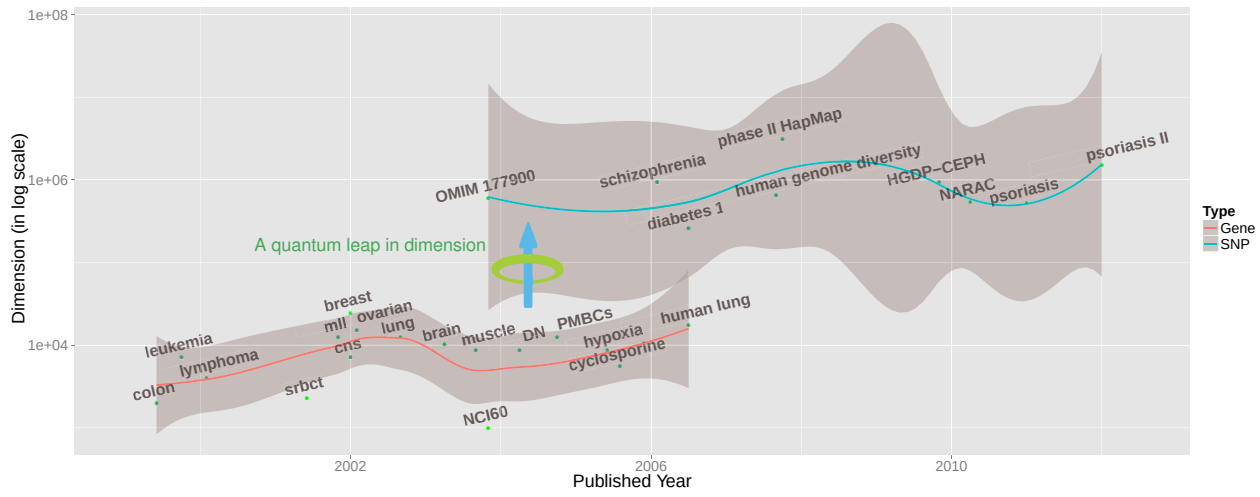


Figure 2.4: An illustration of different biomarker datasets that have been introduced in life science research from the Year 1999 to present. Note that a quantum leap from the use of genes as features (hundred thousands of dimensions) to the choice of SNP as feature descriptors (millions of dimensions) for the identification of relevant biomarkers across a range of BioInformatics or Medical Informatics related datasets can be observed at around Year 2004.

before moving on to *in vitro* study [117, 118]. The rapid advancements in biotechnologies and biodevices, nevertheless, has given researchers the option of using Single-Nucleotide Polymorphism (SNP) (see Figure 2.4) as a new form of feature descriptor that defines the behaviors of genes. Note that this represents a quantum leap from the original thousands of features (genes) to millions of features (SNPs) that one now has to deal with. Similarly for the domain of natural language processing, the feature space is now made up of not only words but phrases or templates that appear in documents, tweet streams and webpages, which also extends to many millions of dimensions.

2.2.3 The Evolution of Feature Dimensions in Data-Centric Research

In this subsection, an analysis on the developments of feature dimensions by focusing on three popular computational intelligence and data analytics repositories is presented, namely, *UC Irvine Machine Learning Repository (UCI)* [119], *UCI KDD Archive (UCI KDD)* [120] and *LIBSVM Database (LIBSVM)* [121], that have transpired in the last two decades. A collection of high dimensional datasets from the three repositories is further

Table 2.2: Characteristics of the datasets collected from 3 popular CI and data analytics repositories. The datasets are grouped according to the domain of application.

Application Domain	Data Name	Dimension	Year
Acoustics	ISOLET	617	1994
game	Chess	36	1989
	Connect-4	42	1995
Image	Letter	16	1991
	Corel	89	1999
Life Science	Soybean	35	1987
	Molecular	58	1990
	Mammals	72	1992
	SPECTF	44	2001
	P53	5,409	2010
Multi-view	Internet AD	1,558	1998
Physics	Spectrometer	102	1988
	H_2O Treat. Plant	38	1993
Sociology	Insurance	86	2000
Text	Bag of words	100,000	2008
	URL	**3,231,961**	2009
Time-series	PEMS-SF	138,672	2011
	Gas Sensor	**1,950,000**	2013
Video	YouTube MVG	**1,000,000**	2013

(a) *UC Irvine Machine Learning Repository*

Application Domain	Data Name	Dimension	Year
Demography	Internet Usage	22	1997
Life Science	E. coli	4,289	2001
Marketing	KDD1998	481	1999
Geology	Forest	54	1998
Text	Microsoft.com	294	1998
	NSF Abstracts	30,779	2003
Times-series	Control Chart	600	1999

(b) *UCI KDD Archive*

Application Domain	Data Name	Dimension	Year
Image	USPS	256	1994
	Gisette	5,000	2003
Life Science	Leukemia	7,129	1999
	Colon-cancer	2,000	1999
	Breast-cancer	7,129	2001
Text	News20	62,061	1995
	Real-sim	20,958	1998
	Sector	55,197	1998
	Rcv1	47,236	2004
	News20.binary	**1,355,191**	2005
	Webspam	**16,609,143**	2006
	SIAM	30,438	2007
	Log1p	**4,272,227**	2009
Education	KDD2010	**29,890,095**	2010

(c) *LIBSVM Database*

considered, whose detailed characteristics including the *year of creation*, *dimension size*, *data name* and *application domain*, are tabulated in Table 2.2.

UCI KDD was originally introduced for use in large-scale data analytics research. This suggests why the datasets available in this archival are higher in dimensions relative to *UCI*. However, as the computational intelligence and data mining communities converge towards Big Data research, there is no longer a need to maintain the *UCI KDD* separately and was since merged with *UCI* from July, 2009. The *UCI* and *LIBSVM* repositories, cover a wide spectrum of real world datasets with various domains, ranging from game `(Chess)`, image `(Corel)`, life science `(Leukemia)`, physics `(Spectrometer)`, text `(Webspam)`, time-series `(Gas Sensor)` to video `(YouTube MVG)` and others.

These data repositories are now becoming *de facto* benchmarks for conducting data analytics studies in many areas of computational intelligence, artificial intelligence, machine intelligence, data mining, soft computing, meta-heuristics and others. The *UCI* for instance, is among one of the top hundred most cited archives[9] in all of computer science related publications, and continues to attract vast interests even today.

To gain understanding on the evolution of feature size (dimension) in data analytics research, the dimensions of the datasets that has been used in the last two decades with respect to the year it was introduced are charted. From Figure 2.5, an exponential increase in dimensionality can be observed across all three popular repositories considered in the early 2000s. For instance, the `News20.binary` is noted to have grown from ten of thousands of features (62,061) in 1995 `(News20)` to one with more than a million in dimension (1,355,191) in just a decade. This is a dramatic rise of more than 20 times. Thus, a simple forecast of the upward trend would reveal that, a feature dimensionality of up to 40 billion features is likely to arise by the year 2020.

2.3 The Challenges and Blessing of Big Dimensionality

The immense growth of feature dimensionality in data analytics has exposed the inadequacies of many computational intelligence methodologies that exist to date. Hence there

[9] `http://archive.ics.uci.edu/ml/about.html`

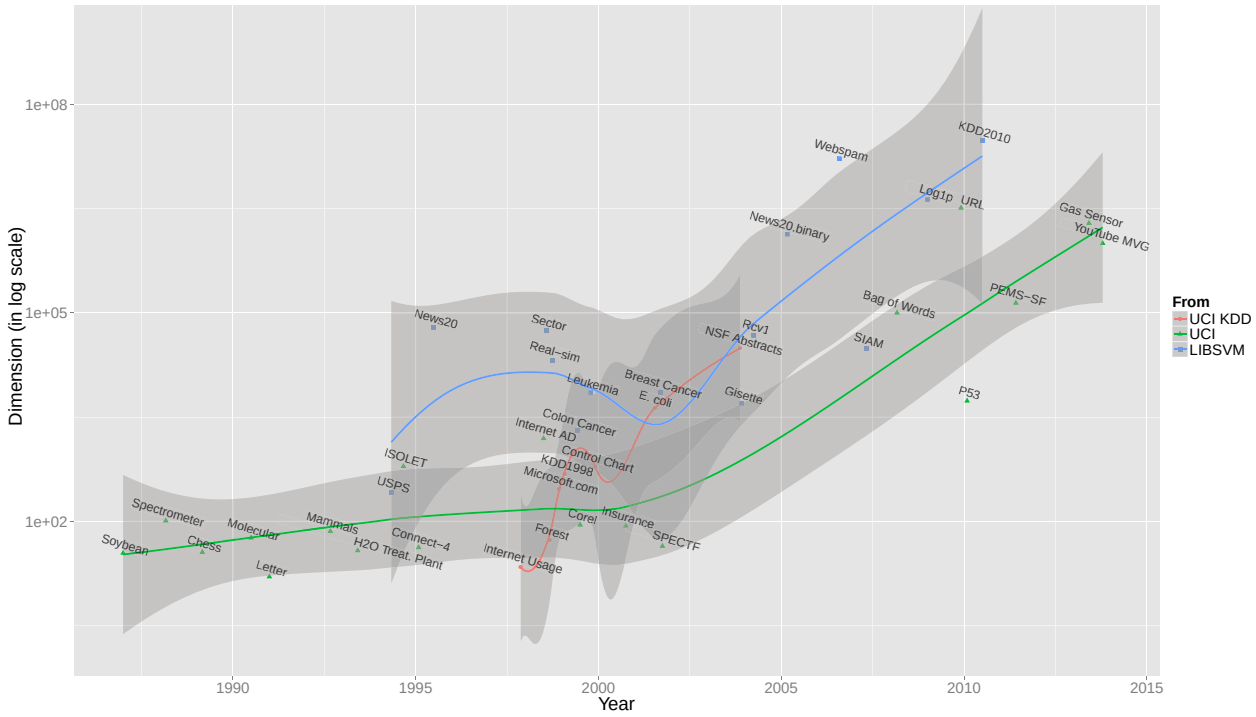


Figure 2.5: An illustration of the evolving feature size (dimension) in data analytics and computational intelligence datasets as introduced in the last two decades.

is an urgent need for the conception of new paradigms and methodologies that can cope with the emerging phenomenon of Big Dimensionality. Correspondingly, how to solicit the key features to concisely represent the data and the prediction model well, while facilitating fast prediction and reduced storage, are among the important tasks of Big Data analytics. To this end, a review on the state-of-the-art feature selection methods is started in this section. Subsequently, the reveal the core emerging challenges of feature selection when facing Big Dimensionality is further proceeded. Lastly, the *"blessing of Big dimensionality"* is also discussed.

2.3.1 Emerging Challenges

From our analysis of the real-world datasets in popular repositories, there is little doubt that Big Dimensionality is setting upon us. The figure of dimensions in research studies on Big Data currently hovers around the million range (10^6). In Table 2.2, it is worth

noting that 7 out of 11 datasets that appeared in the last 8 years have dimensionality in the region of millions.

In this subsection, a discussion on some inadequacies of current computational intelligence methodologies is firstly made, as they were not designed to cope with Big Dimensionality. Consequently, the imperative need for fresh studies on computational intelligence and feature selection paradigms that are proficient in dealing with the explosion of dimensionality and detail some of the core challenges that lies ahead are highlighted.

2.3.1.1 Millions of Dimensions and Beyond

The field of "Big Data" was coined to place attention on the need for new ways in making sense of the unprecedented scale of data that are today becoming ubiquitous. In the same spirit, Big Dimensionality refers to the unprecedented number of features that is scaling to levels which now render existing state-of-the-art computational intelligence approaches inadequate. There is thus a pressing need for new approaches that can cope with this explosion of dimensionality.

In Big Dimensionality, scalability poses as the key challenge to many existing state-of-the-art methods. For instance, the biomarker feature selection problem in life science is cited as the illustrating example. The search for a compact subset of relevant biomarkers [122] from *single-nucleotide polymorphism* (SNP) is known to be critical to biologists in defining the behaviors of genes and their relevance to the disease of interests [117, 118, 123]. In the case of the `psoriasis` SNP dataset, which composes of only 0.5 million features, it took the state-of-the-art SVM-RFE and mRMR biomarker selectors more than a day of computational effort to crunch the data.

To gain further insights into the current state of feature selection research, an analysis on the dimensionality of datasets that have been used in the studies of computational intelligence is conducted, which is summarized in Figure 2.6. Particularly, the focus has been placed on the three flagship transactions and the magazine of the computational intelligence society, namely, the *IEEE Transactions on Evolutionary Computation* (TEC), *IEEE Transactions on Fuzzy Systems* (TFS) and *IEEE Transactions on Neural Networks*

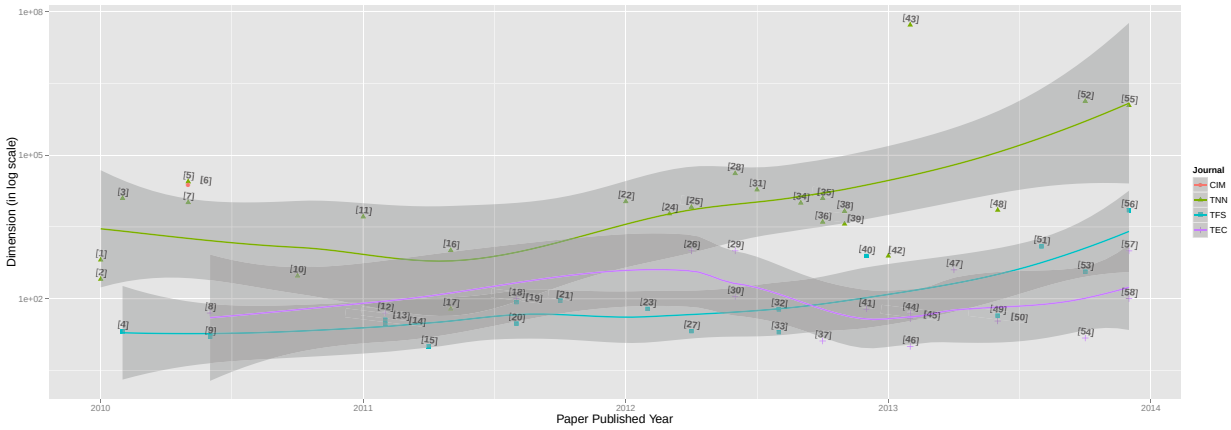


Figure 2.6: Trends of the dataset dimension used in publications that appeared in the flagship journals (TNN, TFS and TEC) and magazine (CIM) of the computational intelligence society from Year 2010 to Year 2013 [1–58].

and Learning Systems (TNN[10]) and *IEEE Computational Intelligence Magazine* (CIM). By contrast, it is evident that as the dimensionality of the datasets continues to rise exponentially in time (see Figure 2.5), the complexity of the feature selection tasks being addressed began to overwhelm the algorithms proposed (see Figure 2.6) to date. Particularly, the dimensionality of the algorithms under-studied (as summarized in Figure 2.6) is significantly lagging behind those being produced (see Figure 2.5). From Figure 2.6, the statistical results further show that a majority (79.3%) of the studies analyzed remain confined to datasets with features that are less than 10,000 in dimensions. Notably, only 5.2% of the studies reported considered real world datasets with features that in the range of millions [43, 52, 55]. In summary, it is becoming clear that the explosion of dimensionality is pushing the capability limits of current computational intelligence algorithms.

2.3.1.2 Handling Trillion Correlations

With millions of features in hand, existing computational intelligence approaches that require the calculations of pairwise correlations in their algorithmic designs will have to cope with computations in the range of trillions[11]. For example, a dataset with millions

[10]Before 2012, this transactions was named as "*IEEE Transactions on Neural Networks*".
[11]The number pairwise correlation computations is a squared of the dimensionality.

of features (10^6) would translate to trillions of pairwise correlations (10^{12}) that need to be computed. However, existing approaches that require the calculations of pairwise correlations in their algorithmic designs (*e.g.*, filters) cannot cope with such datasets elegantly and often scored miserably, since computing at such scale can be intractable. Note that this poses a grand challenge that has never been explicitly addressed in the field of computational intelligence and data mining research.

2.3.2 Blessing of Big Dimensionality

Besides the curse of big dimensionality, in this section, the potential benefits that are attributed by the presence of Big Dimensionality is reviewed, which is less widely noted than the former [124].

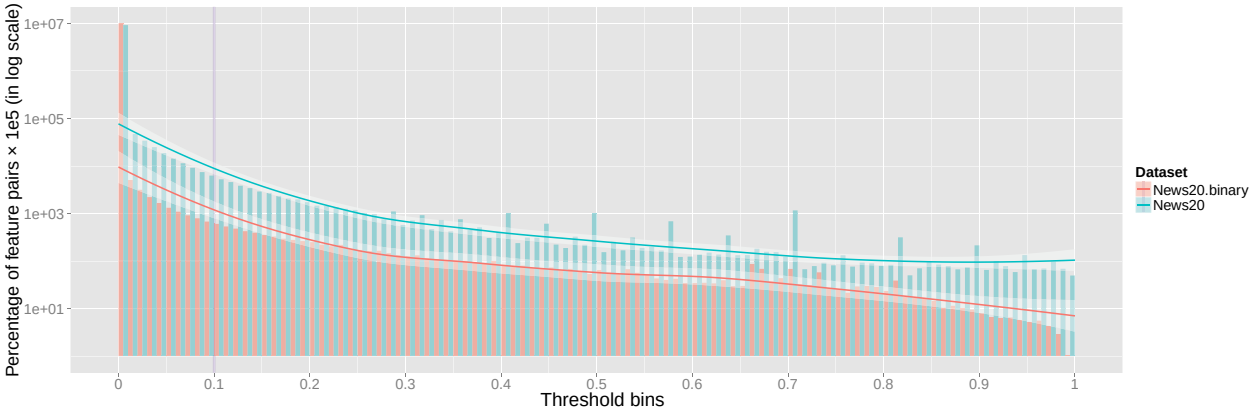


Figure 2.7: Correlation frequencies of feature pairs in the `News20` (62,601 dimensions) and `News20.binary` (1,355,191 dimensions) corpora. Each bar in the figure denotes the percentage of feature pairs (the y-axis) that satisfies the corresponding correlation thresholds (the x-axis).

The results of our experimental study on correlation frequency involving the `News20` (62,601 dimensions) and `News20.binary` (1,355,191 dimensions) corpora are summarized in Figure 2.7. The statistics obtained show that 99.88% and 99.39% of the feature pairs in `News20` and `News20.binary`, respectively, have correlation coefficients that are lower than 0.1. This implies that a majority of the feature pairs are either uncorrelated or the correlated feature pairs are extremely sparse. Moreover, Figure 2.7 displays a downward trend in the number of correlated feature pairs, with increasing correlation threshold.

Further, it can also be observed from the figure that the correlation frequencies of the feature pairs in `News20.binary` are noted to be generally lower than the pairs found in `News20`. This indication of features becoming more sparsely correlated as the dimension scales up clearly showcases a potential blessing of Big Dimensionality that one could leverage upon, since a majority of the uncorrelated feature pairs do not contribute to the correlation matrix.

2.4 Summary

In this chapter, the notion of "Big Dimensionality" has been introduced. In a similar spirit to "Big Data", the term Big Dimensionality has been coined to put attention on the need for new ways in coping with the unprecedented number of features (dimensions) that are scaling to levels that now renders existing computational intelligence approaches inadequate. This survey has revealed the lack of studies on the evolution of data dimensionality in the era of Big Data. In particular, the analysis on three popular data repositories has uncovered an exponential increase in the dimensionality of many datasets that have been produced since early 2000s. In life science research, for instance, a quantum leap from the original thousands of genes (features) to millions of Single-Nucleotide Polymorphism in a short period of time has been observed. And there is much evidence that the upward trend of Big Dimensionality will only continue to follow, as influenced by the rapid advancements in computing and information technologies and the arising myriads of feature descriptors, where a forecast of 40 billion features in dimensions is to be expected by year 2020. Based on our detailed analyses, it has been found that the progress of feature selection methods in the field of computational intelligence are falling very much behind the rapidly rising pace of data dimensionality or Big Dimensionality. Last but not least, the core challenges of feature selection (*curse of Big Dimensionality*) and the potential benefits of dimensionality explosion in feature selection (*blessings of Big Dimensionality*) have been presented and discussed.

Chapter 3

Literature Review

In this chapter, some essential concepts of machine learning that are of interest to the present study, are firstly illustrated. Consequently, the notion of Feature Grouping, as an emerging technique, is introduced, followed by the review of state-of-the-art methods. However, the inference drawn from big dimensional data reveals that there is a lack of capability towards handling feature correlations using existing feature grouping methods. To overcome this issue, by exploiting the blessing of big dimensionality – "sparse correlation", as discussed in Chapter 2, one possible way lies in extending from existing feature selection methods which can efficiently eliminate the non-informative features, while at the same time identifying informative correlated feature pairs from big dimensional data. With this in mind, a review of state-of-the-art feature selection methods is henceforth provided using a traditional taxonomy involving filter, wrapper and embedded methods. Note that, feature correlation and big dimensionality issues are mainly discussed in each method as potential directions for research extension. In addition, a comparison between feature selection and feature extraction is provided in order to substantiate the aim of choosing feature subset from the original feature space.

3.1 Essential Concepts in Machine Learning

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence[1]. Specifically, the early definition from Arthur Samuel in 1959 indicated machine learning as a "field of

[1] https://en.wikipedia.org/wiki/Machine_learning

study that gives computers the ability to learn without being explicitly programmed". However, a widely accepted definition provided by Tom M. Mitchell explains the authentic machine learning of today: "computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [125]. In essence, it is clear that the definitions of machine learning are keeping pace with the times. Beyond the naive interpretation of machine learning, which relates more closely to robotics, this research field is compactly connected with design, analysis, algorithm, implementation and application that are capable of learning from the environment.

Furthermore, an advanced and complete machine learning system is expected to automatically improve its performance on a certain task with gained experience [125]. However, building a comprehensive system is beyond the scope of this thesis. Instead, we focus on filling the gap between feature grouping and big dimensionality on offline (i.e., constant dataset without updating) classification problems. The rest of this section illustrates the concepts that form the crux of this thesis, followed by the core definitions of interest.

3.1.1 Traditional Taxonomy

Machine learning algorithms are often endowed with the nature of learning mechanism – "feedback", as the evidence to change behaviours (i.e., learning or taking lessons). These feedbacks (e.g., "tag", "label" or "class") then serve as meaningful knowledge in differentiating labelled data from unlabelled data. Depending on the type and the availability of the feedback, machine learning approaches are typically classified into three broad categories – Supervised learning, Unsupervised learning and Reinforcement learning [126]:

(i) **Supervised Learning** refers to the task that infers a function/rule from labelled training data. Specifically, the training label plays the role of a teacher, and the resultant function/rule maps the training examples to the training labels.

(ii) **Unsupervised Learning** refers to the task that infers a function/rule to describe the hidden structure from unlabelled data. Since the examples are unlabelled, there

is no feedback to evaluate a potential solution. In addition to discovering the hidden patterns from the data, unsupervised learning can be alternatively performed to study the behavior of features.

(iii) **Reinforcement Learning** refers to the task that interacts with a dynamic environment by performing a certain goal without a teacher as guidance, which is inspired by behaviorist psychology. Note that, this technique is often concerned with how to take actions to maximize the cumulative reward, which is often considered in the game by playing against an opponent.

In this thesis, the emphasis is on offline classification problems, which are always considered in a static environment. As a result, reinforcement learning is beyond the scope of this thesis. Furthermore, in machine learning tasks, classification is considered as the most important component of the CI community, wherein supervised learning acts as the cornerstone. Consequently, in the next two subsections, the key factors in supervised learning are presented. Specifically, the data partitioning (i.e., training set, testing set and V-fold cross validation) is outlined, followed by the details of classification problems.

3.1.2 Training Set, Testing Set and V-fold Cross Validation

The most important goal in machine learning is to discover the predictive relationships from the dataset of interest. Such relationship exploration is usually achieved by learning from one set of observations (i.e., training set) while evaluating on another (i.e., testing set) to examine whether the relationship discovered holds. Formally, a training set is a set of data that is used to discover potentially predictive relationships, while a testing set refers to the set of data that is used to assess the strength and utility of a predictive relationship. Traditional machine learning benchmarks usually have training set and testing set as the evidence for comparisons. However, there are still problems that are represented as a whole (i.e., no specific testing set), or there exists only a small number of available instances in the dataset (especially for big dimensional data). For better performance on these problems, it is necessary to adapt with the idea of resampling, such as "V-fold cross validation".

In V-fold cross validation, a dataset is randomly partitioned into V folds/partitions with nearly equal size, such that the proportion of instances from different classes remains the same in all folds. Subsequently, a single fold of the V folds is retained as a testing set while the remaining (V-1) folds are set together as a training set. The cross-validation process is then repeated V times, and each of the V folds is used only once as testing set. After the training process, the results are usually merged from V folds as an overall outcome by averaging. Notably, since all instances are used for both training and testing, where each instance is used for testing once only, this strategy generally provides stable and robust results, while suffering from overfitting. To be precise, a larger V often leads to smaller bias yet higher probability of overfitting. Further, when V is equal to the number of observations, this extreme case of V-fold cross validation is labelled as Leave-one-out (LOO) cross validation. Note that, LOO is often considered for self-report/self-prediction issues in industries.

To conclude, given the availability of a well split dataset, the subsequent classification task is detailed in the upcoming subsection.

3.1.3 Classification

Generally, classification refers to the process of assigning a given piece of input data (e.g., an instance or example) to one of the given classes/categories. As a result, a classifier must be learnt during the training process. Specifically, the classification algorithm uses a set of examples (i.e., training data) to learn a classifier that is expected to correctly predict the class label of unseen instances (i.e., testing data). The learnt classifier takes the feature values of observations as input and produces the predefined class labels as output. In this thesis, a basic type of classification task is considered in the proposed framework, namely, that of two-class classification or binary classification (e.g., classification for positive and negative).

A typical binary classification example could be considered using patients data. For example, in a hospital, there are some people who suffer from psoriasis (i.e., a type of skin disease) while others who do not. Given a number of people labelled "control" (i.e., healthy people) and others labelled as "case" (i.e., psoriasis patients), the classification algorithm is expected to learn the characteristics of the psoriasis, and the learnt classifier is thus endowed with the ability to process future suspect cases and label them.

3.1.4 Core Definitions

For subsequent ease of exposition, the core definitions are outlined in this subsection. Firstly, m is defined as the dimensionality of the data, while n is the number of training data observations. $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ represents the intact training data, wherein each observation is denoted by $\mathbf{x}_i \in \mathbb{R}^m$, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$. Each vector $\mathbf{x}_i$ is associated with an output label $y_i \in \{\pm 1\}$, and $\mathbf{y}$ is defined as the vector of labels in the training data. Moreover, let $\mathbf{f}_j$ denote a row vector corresponding to the j^{th} feature of all observations in $\mathbf{X}$, thus $\mathbf{X} = [\mathbf{f}_1', \ldots, \mathbf{f}_m']' \in \mathbb{R}^{m \times n}$ holds. Additionally, the element-wise product between two matrices $\mathbf{A}$ and $\mathbf{B}$ is introduced as $\mathbf{A} \odot \mathbf{B}$, and $|\cdot|$ represents the cardinality operator unless specified. Symbols "$\mathbf{0}$" and "$\mathbf{1}$" are the column vectors comprising all zeros and all ones. And for each $\mathbf{f}$, the corresponding mean and standard deviation of the entries in $\mathbf{f}$ are indicated as $\mu_{\mathbf{f}}$ and $\sigma_{\mathbf{f}}$, respectively.

3.2 Feature Grouping

Recently, Feature Grouping has emerged as an intriguing technology that is capable of gathering correlated informative features from the original feature space into various feature groups. Specifically, a feature group may accumulate substantive characteristics of the features, thereby enabling the functional interpretation of the features to the prediction task. Furthermore, by considering each group as a branch, a potential feature structure/network of the dataset can be established by an aggregation of the branches. The knowledge of feature stucture/network futher assists users' understanding/interpretations of the data on hand for further analysis. Furthermore, this technology may also be helpful in grasping the properties of the original feature space of interest.

Consequently, feature grouping has received increasing research interests in the last decade [76, 77, 86, 127, 128]. As pioneers in this research topic, GFlasso [76], OSCAR [128], and ncFGS & ncTFGS [127] have demonstrated promising feature grouping performance on accuracy. In what follows, a review on each of these feature grouping methods is presented.

3.2.1 GFlasso

Lasso is a learning method that performs both selection and regularization in enhancing the prediction accuracy and interpretability. Based on the simultaneity of Lasso, the *graph-guided fused lasso* (GFlasso) is among the early feature grouping approaches and operates by identifying feature groups based on the graph-structure defined over the features that designed for some correlated structures in bioinformatics, such as gene expressions [76]. And the cornerstone, Lasso, is a feature selection approach that designed for linear regression with an ℓ_1 penalty, and it minimizes the sum of squared errors with a bound on the sum of the absolute feature weight [129, 130]. However, Lasso regards the feature as independent trait in the genome-wide association analysis. Taking this cue, GFlasso further employs a sparse regularization over a graph-structure to penalize the differences in feature coefficients β_i and β_j by $|\beta_i - \text{sign}(\rho_{i,j})\beta_j|$. Note that, this is to assess whether features i and j are to be connected with an edge in the graph[2], and then connects/associates $\mathbf{f}_i$ to $\mathbf{f}_j$ when $\rho_{i,j} > 0$. However, the author also suggested that this method can only be applied to the restricted case of correlated continuous-valued outputs. Also the sign function used can create bias in the optimization task [77].

3.2.2 ncFGS & ncTFGS

Recently, Yang *et al.* employs a convex function to penalize the pairwise infinity norm of the connected classification/regression coefficients, while achieving simultaneous feature grouping and selection [127]. Moreover, to alleviate the bias issue, *non-convex (truncated) feature grouping and selection* - (ncFGS & ncTFGS) approaches have been conceived to encourage the sparsity of features (i.e., for feature selection) and equality of absolute values of feature coefficients connected in the graph (i.e., feature grouping). Specifically, for these two methods, they considered non-convex optimization functions to enforce bias alleviation so as to reduce the estimator's variance with the grouped sparsity. The results are shown promising on the bench mark dataset. However, the drawback of this method falls in the necessarily requirement of dependency information, which is usually represented as an undirected graph that is often impractical in many real applications.

[2]The corresponding coefficients β_i, β_j will be similar when $\rho_{i,j} > 0$, but dissimilar when $\rho_{i,j} < 0$, and $\rho_{i,j}$ is the Pearson correlation coefficient between two features $\mathbf{f}_i$ and $\mathbf{f}_j$.

And on the big dimensional data, even if the dependency information is available, the calculation or estimation of the undirected graph can become computational intensive and intractable. Besides, such a method is hardly extended to the problem with directed graphs.

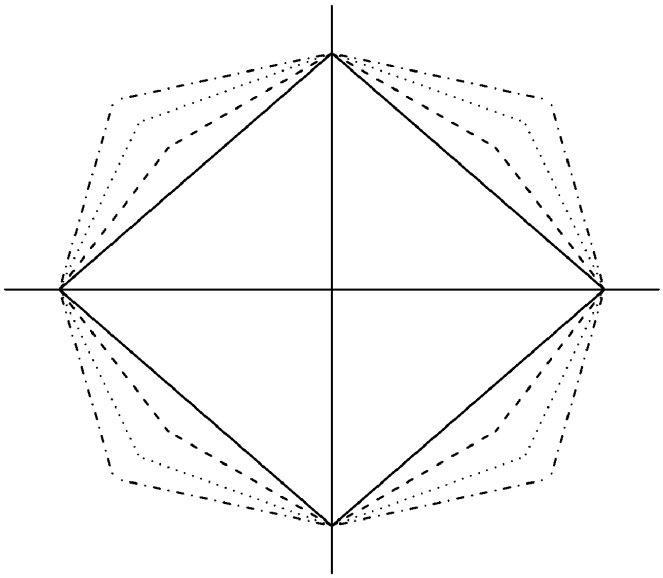


Figure 3.1: Graphical representation of the constraint region in the (β_1, β_2) plane for the OSCAR with variant values of c (The solid line represents the circumstance that when $c = 0$, the method equals to LASSO).

3.2.3 OSCAR

Octagonal shrinkage and clustering algorithm for regression (OSCAR), on the other hand, incorporates the ℓ_∞-penalty as a means to reduce similar feature pairs [128], while the ℓ_1 regularizer is maintained for feature selection purposes. Figure 3.1 illustrates the constraint region of the method for various values of the parameter c. From this figure, the reason for the octagonal term of the method name is obvious, since the shape of the constraint region in two dimensions is exactly an octagon. The OSCAR encourages both sparsity and equality of coefficients to different degrees, depending on the strength of feature correlation, the value of c and the location of the ordinary least squares solution.

However, the optimization is computationally very intensive. Thus, to accelerate the learning process of feature grouping, Zhong *et al.* subsequently introduced an efficient projection step (i.e., iterative group merging) based on the accelerated gradient methods [86].

Note that, althoguth this ℓ_∞-penalty encourages the considered feature coefficients to be similar, thus forming feature groups (e.g., features i and j will be in one group if $|\,|\beta_i| - |\beta_j|\,| < \epsilon$), the drawback, nevertheless, falls on the assumption that all features are connected, which is obviously unsuitable for many applications, especially when the number of features is getting large.

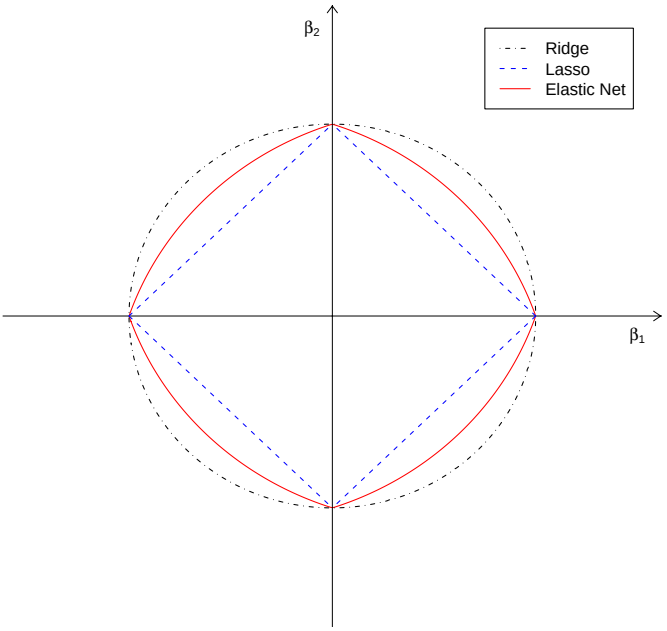


Figure 3.2: 2-dimensional contour plots: singularities at the vertices and the edges are strictly convex; also the strength of convexity varies w.r.t. α (the penalty of elastic net is with $\alpha = 0.5$).

Other works include the Elastic-Net [131] and group Lasso [132], wherein the former uses the hybridization of ℓ_1 and ℓ_2 regularizers to gather strongly correlated features into groups when dealing with high dimensional problems, as illustrated in Figure 3.2; the latter, however, introduces an extension of the lasso penalty, which is deemed as an intermediate between ℓ_1 and ℓ_2 penalty, so as to favor robust features. However, these

methods require the feature groups to be given as priori, which means that such methods aim at learning sparse structure of a given model, rather than building the structure from scratch. Consequently, for big dimensional data, these methods face scalability problems. Thus, in contrast to this refining/pruning scheme (i.e., seeking the feature groups as sub-network from the overall feature-network), building the feature structure from scratch is preferred in dealing with big dimensional data.

3.2.4 Conclusion on Feature Grouping Facing with Big Dimensionality

In spite of the increasing efforts that focus on identifying intrinsic feature groups, existing feature grouping strategies have met with limited success on big dimensional data. The key factor that is responsible for this phenomenon is the inevitability of computing the extremely large covariance/correlation matrix on big dimensional data, which is computationally intractable for many state-of-the-art methods. Notably, a dataset with millions of features translates to trillions of correlations to be computed, thus even a matrix approximation method can easily fail to be efficient. In contrast to previous works, this thesis aims to provide a way of exploiting the presence of sparse correlations for the efficient identifications of informative and correlated feature groups from big dimensional datasets. And one of the key objectives is to avoid the calculation of a full covariance/correlation matrix. In particular, the idea is to first select a feature subset that contains the most informative features, and a grouping/clustering strategy is subsequently performed among the top features to acquire the correlated features that contribute to the feature structure. Based on this approach, the manner of selecting the informative features plays a key role. Accordingly, in the next section, the technique of seeking key features is reviewed in detail, along with state-of-the-art methods based on the traditional taxonomy.

3.3 Feature Selection

In comparison to the recently proposed conceptof feature grouping, feature selection is considered as a more mature and well-established branch of machine learning in the literature. Over the decades, a vast variety of feature selection methods have been proposed

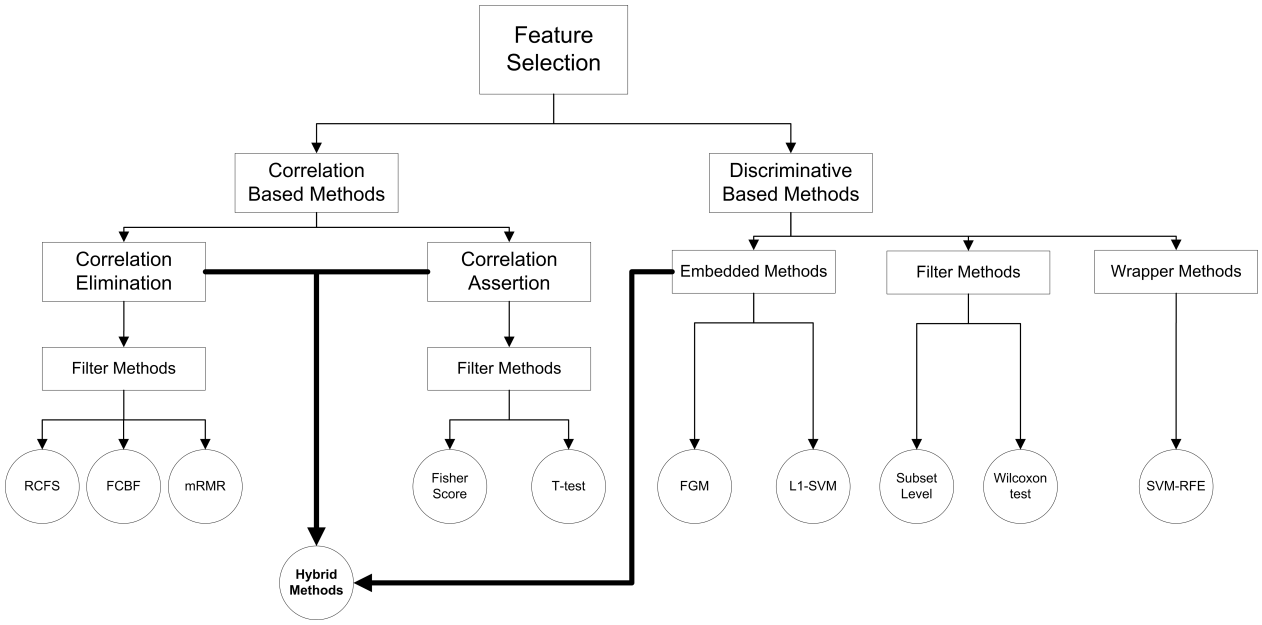


Figure 3.3: Detailed categorization of considered feature selection methods with respect to correlation consideration. (Hybrid Method is the desired method which takes the advantages of the state-of-the-art methods.)

for handling different learning scenarios. Hence, many principles for the categorization of feature selection methods have emerged to date. As this research focuses on the potentials of correlated features on feature selection problems involving big dimensions, a detailed taxonomy based on correlation and discriminative feature selection methods is presented here, as depicted in Figure 3.3. The correlation-based category consists mainly of the filter methods, which have served as useful tools for data analysis. In the discriminative based category, all of the aforementioned three core themes of approaches, namely, filter, wrapper and embedded methods, have been studied. As a counterpart of embedded methods, which involve the ideas of both correlation elimination and correlation assertion, there exists "Hybrid Methods" that take advantage of diverse benefits from multiple methods that follow the research interest. In the remainder of this subsection, the benefits, drawbacks, and examples of state-of-the-art filter, wrapper and embedded methods are discussed. Notably, two perspectives are mainly considered for the feature grouping extension: feature correlation consideration and scalability towards big dimensional data.

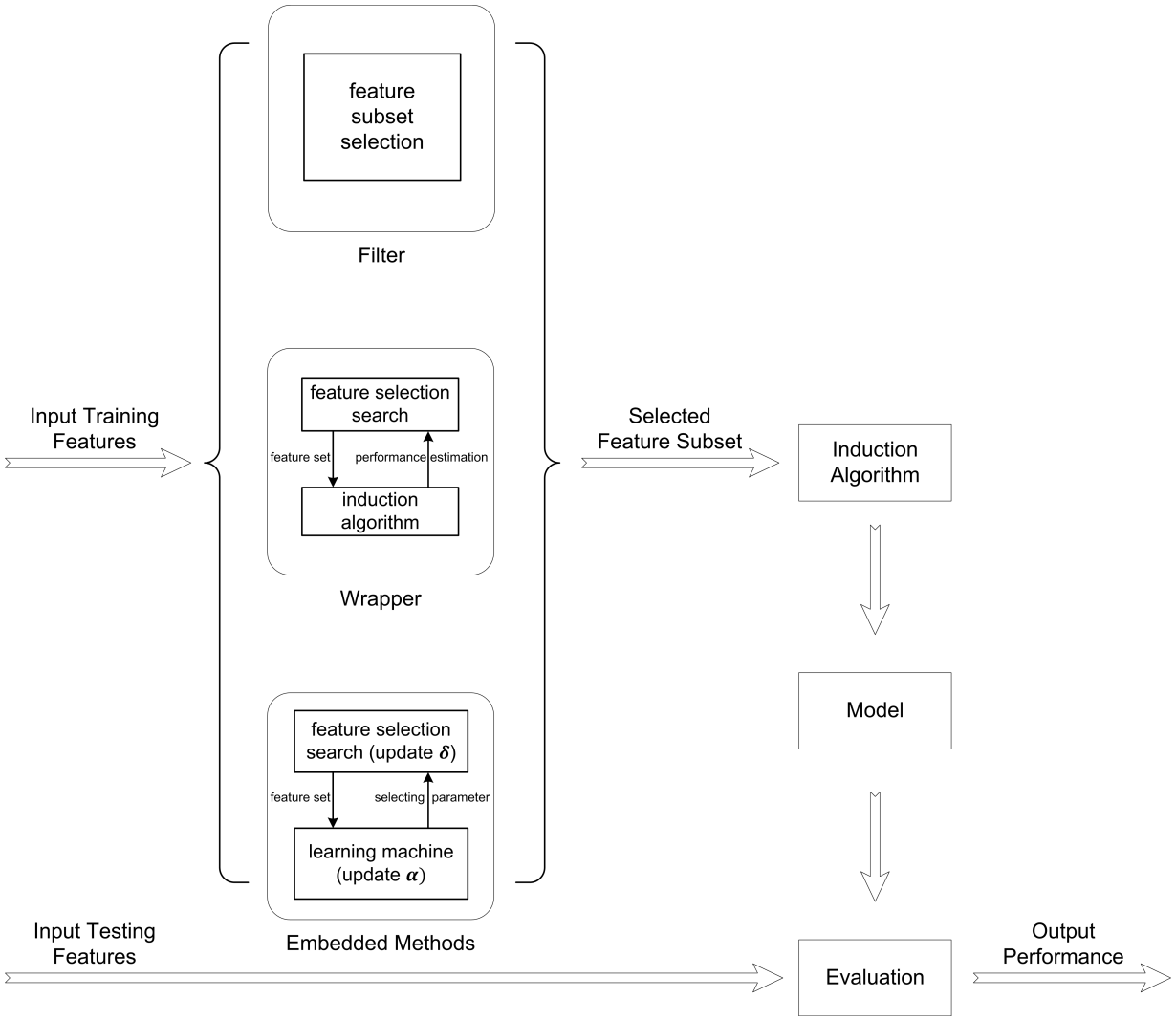


Figure 3.4: Operating Principles of Filters, Wrappers and Embedded Methods.

3.3.1 The Traditional Taxonomy of Feature Selection

The task of dimension reduction with consistent feature space has been regarded as desirable in many applications aforementioned. In the recent decades, a plethora of feature selection methods have been developed to identify the important feature subset. In general, the methods of feature selection have been categorized into three core themes based on their operating principles [79]: **filter methods** [133, 134], **wrapper methods** [118, 135] and **embedded methods** [136–138], as depicted in Figure 3.4. Specifically, **filter methods** select informative features based on their individual discriminative

power such as the correlation criterion. The benefits of filter methods lie in their low computational requirements or high efficiency in leading to a small number of core subset features. The drawback, however, is that it may not identify the optimal feature subset suitable for the predictive model of interest, and is generally incapable of handling very high dimensional problems. On the contrary, since **wrapper methods**, such as gene selection based on Support Vector Machine (SVM) and Recursive Feature Elimination (SVM-RFE) [118], select the discriminative features solely based on the inductive learning rule, they typically exhibit more precise performance in prediction, but at the expense of a higher computational cost incurred, especially on large scale and very high dimensional problems. More recently, there has been increasing interests on **embedded methods**, which refer to approaches that directly incorporate the feature selection scheme within the learning machine, for instance, by optimizing some regularized risk function $g(\boldsymbol{\alpha}, \boldsymbol{\delta})$ with respect to two sets of parameters: parameter $\boldsymbol{\alpha}$ of the learning machine, and parameter $\boldsymbol{\delta} \in \{0, 1\}^m$ [79] to control feature sparsity. As such they are usually more efficient than wrapper methods. An established embedded method is the state-of-the-art ℓ_1-regularized Support Vector Machine (L1-SVM) [136].

3.3.2 Filter Methods

Filter methods select features on the basis of their relevance or discriminant powers with regard to the targeted classes. Simple methods based on mutual information and statistical tests (t-test [139], F-test [140]) have been proven to be effective. Consequently, the selected features have better generalization properties, which is to say that the selected features from training data generalize well to new unseen data.

In filter methods, two popular filter metrics for classification problems are feature correlation and mutual information, although neither is true metric or distance measure in the mathematical sense. Since they fail to obey the triangle inequality and thus do not compute any actual "distance", they should rather be regarded as "scores". That is how the notion of "feature score" that has been introduced in filter methods, and these feature scores aim to measure the similarity between a candidate feature (or a set of features) and the desired output category. There are, however, true metrics that are a simple function of the mutual information. Based on these correlation measures,

several remarkable methods have attempted to reduce the redundancy among the selected features.

3.3.2.1 The Basic Idea of Filters

Identifying the most informative feature subset from the original feature space is of a significant but challenging problem, which can be formulated as a combinatorial optimization problem [141]. Given a feature index set $\mathcal{F} = \{1, 2, ..., m\}$, actually "feature subset identification" is to find a feature index subset $\mathcal{F}^* \subseteq \mathcal{F}$ by maximizing the objective function $f : \Pi \to R$, such that

$$\mathcal{F}^* = \arg\max_{\mathcal{F}_s \in \Pi} f(\mathcal{F}_s) \tag{3.1}$$

where Π is the space for all possible feature subsets in $\mathcal{F}$, and $\mathcal{F}_s$ denotes a subset of Π. Moreover, it is worth mentioning that the optimal subset of a dataset shall not be necessarily unique. Take the consideration of redundancy reduction aside, obviously, an intuitive and convenient feature selection method is to rank the features via feature importance score, hence the features which take the largest scores should be considered for constructing the eventual feature subset, and such approach is typically followed in filter method [7, 133, 134].

3.3.2.2 Subset-level Score

Traditionally, filter method always consider the utility of each individual feature. Formally, these filters intend to generate the feature score based on the feature subset, which is to optimize the following maximization problem

$$\mathcal{F}^* = \arg\max_{\mathcal{F}_s \subset \mathcal{F}, j \in \mathcal{F}_s} c_j, \tag{3.2}$$

where c_j is regarded as the feature score for measuring the predictive ability (correlation with the class) of j^{th} feature, i.e., $\mathrm{SU}(\mathbf{f}_j|\mathbf{y})$, representing the Symmetrical Uncertainty [142], which is a form of non-linear correlation measure. Since these filters pay less consideration on the interactions among features, for example, in applications such as microarray data analysis [68, 80], they often face the problem of **suboptimality** (local optima).

To tackle this problem, instead of calculating feature-level score, the idea of directly measuring the contributions of a feature subset is introduced and further a general graph-based feature selection framework under trace ratio criterion is also studied in [141]. As graph depicts the relationship among data in a natural and effective way, weighted undirected graphs can be generated to encode the within-class and between-class information of the data, and the corresponding matrices are A_w and A_b, respectively. Generally, $(A_w)_{ij}$ ($(A_b)_{ij}$) is a relatively larger value if instance $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same (different) class or are close (distant) to each other, while smaller otherwise. In detail, differing from (3.2), **subset-level score** method is proposed to measure the importance of a feature subset, where $\sum_{ij} \parallel y_i - y_j \parallel^2 (A_w)_{ij}$ should be minimized and $\sum_{ij} \parallel y_i - y_j \parallel^2 (A_b)_{ij}$ be maximized [141], such that

$$\mathcal{F}^* = \underset{\mathcal{F}_s \subset \mathcal{F}, i,j \in \mathcal{F}_s}{\arg\max} \frac{\sum_{ij} \parallel y_i - y_j \parallel^2 (A_b)_{ij}}{\sum_{ij} \parallel y_i - y_j \parallel^2 (A_w)_{ij}}. \tag{3.3}$$

With the above formulation, a feature subset can thus be identified such that the subset-level score is maximized. Moreover, this maximization problem can be globally solved through an iterative algorithm termed as the subset-level score method (i.e., S-FS and S-LS in [141]). In general, subset-level score method can obtain better feature subset than simple filter methods do. However, from (3.3), it is easy to observe that, the subset-level score only considers a simple combination of features, hence suboptimal solution to the output labels still remained. In addition, since this method scales with $O(n^2m)$, it is inefficient for dealing with very high dimensional problems.

3.3.2.3 Minimum Redundancy Maximum Relevance

Mutual Information, as defined in Equation (4.3), is a nonlinear correlation metrics. Let X and Y be two continuous random variables with joint probability density function (pdf) $p(x, y)$, and marginal pdfs $p(x)$ and $p(y)$, respectively. Clearly, as Figure 3.5 illustrates, Battiti [143] defined the feature reduction problem as the process of selecting the most relevant features from the initial set of features based on mutual information to identify the relevancy and redundancy among features, via a proposed greedy selection scheme.

As a family of Battiti's Mutual Information Feature Selection (MIFS) [143], one notable redundancy reduction feature selection method is the Minimum Redundancy Maximum Relevance (mRMR) [134], which selects the most correlated features that contribute

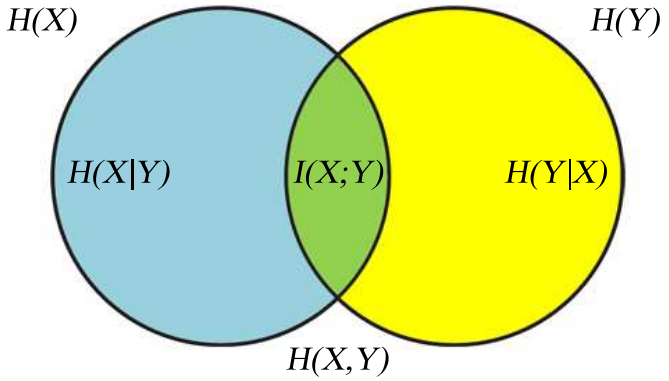


Figure 3.5: Illustration of various information theoretic quantities. (Joint $H(X, Y)$, individual $(H(X), H(Y))$, and conditional entropies for a pair of correlated subsystems X, Y with mutual information $I(X; Y)$).

to the labels such that they are mutually far apart from each other by maximizing the dependency between the joint distribution of the selected features and the output labels. Let S denote the features subset that one is seeking and Ω the pool of all candidate features. Thus, the relevance (dependency) of a feature set S for the class $\mathbf{y}$ is defined by the average value of all mutual information between the individual feature $\mathbf{f}_i$ and the class as follows,

$$D(S, \mathbf{y}) = \max_{S \subset \Omega} \frac{1}{|S|} \sum_{\mathbf{f}_i \in S} I(\mathbf{f}_i; \mathbf{y}). \tag{3.4}$$

Also, the redundancy of all features in the set S is defined as the average mutual information between the individual feature $\mathbf{f}_i$ and $\mathbf{f}_j$

$$R(S) = \min_{S \subset \Omega} \frac{1}{|S|^2} \sum_{\mathbf{f}_i, \mathbf{f}_j \in S} I(\mathbf{f}_i; \mathbf{f}_j). \tag{3.5}$$

Hence the mRMR criterion is a combination of the above two measures and is defined as follows,

$$\max_S [D(S, \mathbf{y}) - R(S)] = \max_S [\frac{1}{|S|} \sum_{\mathbf{f}_i \in S} I(\mathbf{f}_i; \mathbf{y}) - \frac{1}{|S|^2} \sum_{\mathbf{f}_i, \mathbf{f}_j \in S} I(\mathbf{f}_i; \mathbf{f}_j)]. \tag{3.6}$$

This incremental selection scheme of mRMR avoids the multivariate density estimation when maximizing dependency, while the generality allows mRMR to be compatible

to many others. Besides, it can be effectively hybridized with other feature selection methods, such as the wrappers, for seeking a very compact subset from the candidate features. Moreover, it may be shown that the mRMR is an approximation of the theoretically optimal maximum-dependency feature selection that maximizes the mutual information between the joint distribution of the selected features and the classification variable. Since mRMR considers the combinatorial problem as a series of smaller scale sub-problems, each of which involves only two variables, the estimation of joint probabilities is considered more robust.

However, since the condition of mRMR is equivalent to the maximal dependency condition for first-order feature selection, this method also suffers from the computational speed. Further, in some situations, the algorithm can underestimate the usefulness of features, especially when it has little way of measuring the interactions between features. This can lead to poor performance when the features are independently redundant [144], but contributing when combined with other features (a pathological case is found when the class is a parity function of the features).

3.3.2.4 Fast Correlation-Based Filter

Another feature selection method is the Fast Correlation-Based Filter (FCBF) [133], where feature importance and feature correlations are assessed by means of the SU measure. It is composed of two core steps, namely, selecting a subset of relevant features, and selecting predominant features from relevant ones, and detailed as follows:

(i) Firstly, the algorithm calculates the SU value for each feature, the relevant features (relevant to the output labels) are recorded into a list based on some predefined thresholds, and then sorted in a descending order, according to their SU values.

(ii) The ordered list is further processed to select the predominant features based on some elegantly designed intuitive rules [133].

(iii) Based on the correlation trick of Markov Blanket [145], which is discussed in [135], such that a feature $\mathbf{f}_j$ which has already been determined to be a predominant feature can be used to filter out other features, for which $\mathbf{f}_j$ forms an approximate Markov blanket. Since the feature with the highest class-correlation does not have any approximate Markov blanket, it must be one of the predominant features.

FCBF can efficiently achieve high degree of dimensionality reduction and enhance or maintain the predictive accuracy with selected features. It is worthy to emphasize that feature subsets selected by FCBF are decoupled from the choice of learning algorithms. In other words, FCBF does not directly aim to increase the accuracy of a particular learning algorithm, like what wrapper methods do. In order to achieve better accuracy within affordable time, a wrapper algorithm based on an intended learning algorithm can be applied to the significantly reduced subset obtained by FCBF.

Major computation of the algorithm involves SU values for correlation measure, which has a linear complexity in terms of the number of instances in a dataset. The algorithm has a linear complexity $\mathrm{O}(m)$ for identifying the relevant features; a best-case complexity $\mathrm{O}(m)$ to determine predominant features from relevant ones when only one feature is selected and all of the rest of the features are removed, and a worse-case complexity $\mathrm{O}(m^2)$ when all features are selected. However, in general cases when k ($1 < k < m$) features are selected, the number of evaluations performed by FCBF will typically be much less than the number of evaluations performed by greedy sequential search, since features removed in each round are not considered in the next round. This makes FCBF substantially faster than algorithms of subset evaluation based on greedy sequential search. The more features removed in an earlier round, the faster FCBF is. Moreover, selecting a subset of relevant features in the first step can further improve the efficiency. However, the method is prone to introduce a number of noisy features, which makes the method less robust.

3.3.2.5 Redundancy Constrained Feature Selection

Recently, Zhou *et al.* proposed a work named Redundancy Constrained Feature Selection (RCFS) [7]. The essence of this method is to first perform feature clustering based on some distance measures (e.g., $1 - |\rho(\mathbf{f}_j, \mathbf{f}_k)|$). Hence there exists a pretty high possibility that the correlated features are grouped into several clusters. After that, some more significant features are then identified from each cluster. In the second step, a feature subset is further determined from the selected features in each cluster under some graph based feature selection criteria [141] to capture the global or the local intrinsic structures of the dataset. This strategy, however, can be heavily sensitive to the choice of graph Laplacian matrices used. For example, the Laplacian score is usually constructed using K

nearest neighbors (KNNs). In practice, when faced with very high dimensional problems, KNNs will no longer be very meaningful since the KNNs of a certain feature might be very far away from each other in reality, due to the effect of the "Curse of Dimensionality". Besides, the high computational cost of feature clustering on the high dimensional data and graph based methods (taking $O(n^2m)$) make this approach less attractive on large scale data.

Recently, Zhao *et al.* proposed a framework to unify different criteria for removing feature redundancies [81] (i.e., minimize the **redundancy rate**). Nevertheless, existing methods have remained to focus on reducing these redundancies rather than how the features correlate together. However, to date, the discovery of correlated yet informative features has been relatively unexplored although this term had been proposed for many years.

3.3.3 Wrapper Method

Wrapper methods, on the contrary, evaluate feature subsets with respect to the classification algorithm in mind, and hence measures the effectiveness according to classification accuracy [146]. Thus, feature subsets are generated based on some search strategy, and the feature subset which leads to the best correct classification rate is identified. Among the algorithms widely used, Genetic Algorithm (GA) [135] and Sequential Forward Selection (SFS) methods are among the prominent examples. The computational complexity is higher than filter methods, but the selected subsets are generally more effective for the inductive machine of interest, even if they remain sub-optimal. Wrapper methods wrap feature selection around a specific prediction method; the prediction method's estimated accuracy directly judges a feature's contributions. Thus one can often obtain a set containing small number of features that offers high accuracy since the features' characteristics match well with the learning method of interest. Typically, wrapper methods require extensively more computation to search the best features than filter methods.

3.3.3.1 SVM-RFE

Support Vector Machine (SVM) has been established to operate well on large-scale and high-dimensional classification problems [147]. Naturally, it is expected that numerous

1 Initially set the index subset of surviving features $\mathbf{s} = \{1, 2, \cdots, m\}$, and feature ranked list $\mathbf{r} = \varnothing$.
2 **repeat**
3 Train a learning machine f on the current subset of features based on SVM, by minimizing a risk functional $J(f)$.
4 Compute the weight vector of current dimension length of $\mathbf{s}$.
5 For each remaining feature, estimate the change in $J(f)$ without retraining on f, using above weight vector. (e.g., change of removing this remaining feature)
6 Select some features $\mathbf{t}$ which will lead to an improving or least degrading on $J(f)$ and update the feature ranked list $\mathbf{r} = \mathbf{r} \cup \mathbf{t}$.
7 Remove the features of $\mathbf{t}$ from $\mathbf{s}$.
8 **until** $\mathbf{s} = \varnothing$;

Algorithm 1: SVM-RFE

schemes on using SVM for feature selection and classification on high dimensional data have been proposed [74, 148]. For example, Support Vector Machine based on Recursive Feature Elimination (SVM-RFE) for gene selection has been introduced in [118].

The outline of SVM-RFE is given in Algorithm 1, where SVM-RFE firstly searches through all the features individually, leading to a long searching process, especially on large-scale and very high dimensional data. By introducing a scaling vector to the input features, the feature selection problem can be formulated as a joint optimization problem of SVM training. SVM-RFE is shown to attain nested subsets of input features that exhibit state-of-the-art performance on gene selection involving microarray data. Furthermore, Rakotomamonjy showed the effectiveness of feature selection based on various SVM-based criteria [149].

As the method selects the discriminative features w.r.t. the inductive learning rule, they typically exhibit more precise performance in prediction, but at the expense of lower computational efficiency on large scale and very high dimensional problems.

3.3.4 Embedded Methods

Embedded methods differ from other feature selection methods in that the feature search mechanism is directly embedded and becomes part of the classifier model used. Particularly, filter methods do not incorporate the learning model, while wrapper methods

involve a learning machine that measures the quality of subsets of features without incorporating prior knowledge about the specific structure of the classification or regression function, hence any form of learning machine can be used. In contrast to filter and wrapper approaches, in embedded methods, the learning machine and the feature selection scheme are tightly coupled, i.e., the structure of the class of functions under consideration plays a crucial role.

The way of generating embedded methods can be regarded as good inspiration to design new hybrid feature selection techniques for specific algorithms, since it can find a function that represents the prior knowledge about what a good model is. Moreover, owing to the fact that this approach will directly optimize some regularized risk functions $g(\boldsymbol{\alpha}, \boldsymbol{\delta})$ with respect to two sets of parameters, it can use early stopping (validation set) or some special stopping criteria to stop and select the subset of features based on optimizing alternatively according to $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$.

3.3.4.1 ℓ_1-Sparse Support Vector Machines

The traditional ℓ_1-norm SVM, as proposed by Bradley and Mangasarian in [150], can be generalized to a general ℓ_1 Support Vector Machines (L1-SVM). It has been proven that solving the corresponding optimization problem gives a smaller error penalty and enlarges the margin between two support vector hyper-planes, thus possibly giving better generalization capability than traditional SVM. A linear decision hyperplane $f(x) = \mathbf{w}'\mathbf{x}$ is learnt by minimizing the following structural risk functional:

$$\min_{\mathbf{w}} \quad \Omega(\mathbf{w}) + C \sum_{i=1}^{n} l(-y_i \mathbf{w}'\mathbf{x}_i), \tag{3.7}$$

where $\mathbf{w} \in \mathbb{R}^m$ is the weight vector of the decision hyperplane, and $\Omega(\mathbf{w})$ is the regularizer that defines the characteristic of $\mathbf{w}$ (e.g., sparsity). $l(\cdot)$ denotes a convex loss function[3], and $C > 0$ is a regularization parameter that trades off between model complexity and the fitness of the decision hyperplane. In traditional SVM, the regularizer $\Omega(\mathbf{w})$ is set to $\frac{1}{2}\|\mathbf{w}\|_2^2$, which is usually non-sparse and so is the learned weight vector $\mathbf{w}$. Furthermore, to obtain a sparse decision rule for SVM, one of the most widely used approaches is to

[3]In the context of SVM, the maximum margin decision hyperplane is usually learned based on either hinge loss or square hinge loss.

introduce ℓ_1-regularizer on the loss function [150–152], resulting in the following ℓ_1-norm SVM problem,

$$\min_{\mathbf{w}} \quad \|\mathbf{w}\|_1 + C\sum_{i=1}^{n} l(-y_i \mathbf{w}'\mathbf{x}_i), \tag{3.8}$$

Although the level of sparsity in the decision function of (3.8) can be controlled by varying C, due to the regularization term $\|\mathbf{w}\|_1$, the bias problem inevitably exists in ℓ_1-regularized optimization problems [153]. Specifically, for a feature $\mathbf{f}_j$, the corresponding weight w_j would be very large if it is highly informative for the classification task and important to reduce the empirical loss. However, the value will be suppressed due to the minimization of $\|\mathbf{w}\|_1$ in (3.8), with the scaling bias.

In addition, due to the scale variation of $\mathbf{w}$, it is hard to control the sparsity while regulating the decision function. Finally, the ℓ_1-regularization can be inefficient when handling large-scale and very high dimensional problems, especially on the datasets which are too large to be loaded into the memory.

3.3.4.2 Feature Generating Machine

Recently, Tan *et al.* proposed an embedded novel feature selection work, named Feature Generating Machine (FGM), has been shown with cracking scalability to non-monotonic feature selection on both large scale and very high dimensional datasets [137]. It uses an SVM to learn a sparse input variable subset for classification, while minimizing a convex relaxation (by introducing an additional variable $\theta \in \mathbb{R}$, as indicated by the following formulation) of the constrained optimization. However, it is improbable to be solved since there are infinite number of quadratic inequality constraints, which is in the form of the Semi-Infinite Programming (SIP) problem [154]. Fortunately, the *Cutting Plane Algorithm* [154, 155] is an efficient optimization scheme for solving this problem. Note that, the cutting plane strategy is employed to efficiently find the sparse weights, by iteratively solving the convex relaxation problem as follows,

$$\max_{\alpha \in \mathcal{A}, \theta} -\theta \quad : \quad \theta \geq -S(\alpha, d^t), \quad \forall d^t \in \mathcal{D}. \tag{3.9}$$

This method has been shown to produce state-of-the-art results in terms of sparsity and generalization performance, especially for very high dimensional datasets, and the

optimal solution can be obtained without any numeric optimization solver. Moreover, the framework introduces a feature indicator vector $\mathbf{d} \in \{0,1\}^m$ into the traditional SVM to make it an ℓ_0-norm Sparse SVM (SSVM) model, the resultant formulation of SVM is then transformed to a Mixed Integer Programming (MIP) problem. Furthermore, the term "Feature Generation" refers to solving a convex relaxation of the MIP problem. The drawback of FGM lies in that it is only formulated for the 0-1 loss function, and the feature correlation is not considered well in the method, the resultant feature subset always maintain highly correlated yet informative features, leading to the possible exclusion of crucial features.

However, since FGM can perform a fast search and report a bunch of highly informative features of big dimensional data. The feature grouping should hence be performed among these informative features. This inspires us in seeking constraint that corporate feature dependency with feature information, hence in generating an up to date feature grouping method.

3.3.5 Feature Selection vs. Feature Extraction

As mentioned in the introduction, Dimension Reduction is considered as a well-established remedy for addressing the "Curse of Dimensionality". Generally, Feature Extraction is considered as another way of reducing the number of features under consideration. By definition, feature extraction transforms the data from a high dimensional space to a space spanned by fewer dimensions. While the data transformation can be linear, as for the case of Principal Component Analysis (PCA) [156, 157], yet many other nonlinear techniques also exist [158, 159].

Although feature extraction is recognized as a flexible (deformable) strategy, it attempts to construct some new joint features by combining individual original features based on their dependency structure [160], resulting in inconsistent feature space (i.e., the extracted features are not invariant under transformation), which has been regarded as a drawback on some tasks (i.e.., the extracted high level feature could not be clearly represented). Feature selection, on the other hand, focuses on finding a small feature subset among the **original feature vectors**. Note that by performing feature selection, the original feature space is still maintained. Further, this property is also of relevance to

feature grouping, where the grouped features are expected to eventually form a network. Hence, in this thesis, feature extraction is not explicitly studied as a matter of research interest.

3.3.6 Summary of Feature Selection As A Cue for Advanced Feature Grouping

In this section, a review of state-of-the-art feature selection methods, namely, filter, wrapper and embedded approaches, is presented. To summarize, filter methods based on correlation consideration such as mRMR, FCBF and RCFS, return good feature subsets with low redundancy and produces reliable accuracy performance rapidly. Filter methods are however less attractive on big dimensional datasets, since their use of feature correlation computations or class dependency calculations do not scale well with increasing dimensions. Neither the present wrapper nor embedded methods consider the correlation among the features in the search. Both have been shown also to generate good prediction accuracy, but embedded approaches have been demonstrated to cope better on very high dimensions. Nevertheless, both wrapper and embedded methods are established to incur extensive computations when searching for the best feature ranking according to the inductive learning rules used.

In the past decade, a variety of feature selection approaches have focused on introducing new schemes to reduce the redundancy among the selected features. The notion of feature redundancy is usually measured by means of pairwise feature correlation under some correlation metrics. The major motivation has been to find the optimal or minimized feature subset corresponding to the output labels. Thus, when a feature is selected, other features that are highly related to the corresponding selected feature are typically rejected so as to minimize feature redundancies. However, as previously stated, **feature grouping, together with efficient feature selection on big dimensional datasets**, are of particular interest to the present study in order to fill the voids in the existing research landscape. Thus, as described previously, the proposed strategy not only provides a feature index subset, as is the case for traditional methods, but additionally involves a comprehensive study on latent feature group structures that exist within the original feature space. With this background, the pursuit of the desired goal is presented in the chapter that follows.

Chapter 4

Group Discovery Machine

4.1 Introduction

Feature correlation is among one of the most commonly used criteria of feature selection tasks in machine learning and data mining. While some researchers have focused on minimizing the correlations among features in the identified feature subset [78, 161, 162], others have exploited the mechanism of feature correlations via feature groups that capture new salient characteristics of the data [76, 127, 163]. These feature groups then serve as cues that could assist the human user in further analysis of the data. From a survey of the literature, feature correlation has been widely established as an important criterion for the identification of relevant, irrelevant, redundant and/or noisy features in learning and prediction tasks. It has received tremendous attentions over the past decades since datasets comprising a large number of features are now becoming ubiquitous [80, 161–163].

From a survey of the literature, today, modern databases with "Big Dimensionality" (i.e., millions of dimensions and above, as discussed in [59]) are becoming evident and such phenomenon will continue to be a growing trend. As the dimensionality of datasets continues to push the capability limits of the algorithms, it is becoming clear that the complexity of the feature grouping and selection tasks being addressed began to overwhelm the algorithms available, i.e., due to the exponential increase in data dimension. In particular, existing approaches that require the calculations of pairwise correlations in their algorithmic designs cannot cope well with such high dimensional datasets elegantly and often scored miserably, since computing the full correlation/covariance matrix (i.e.,

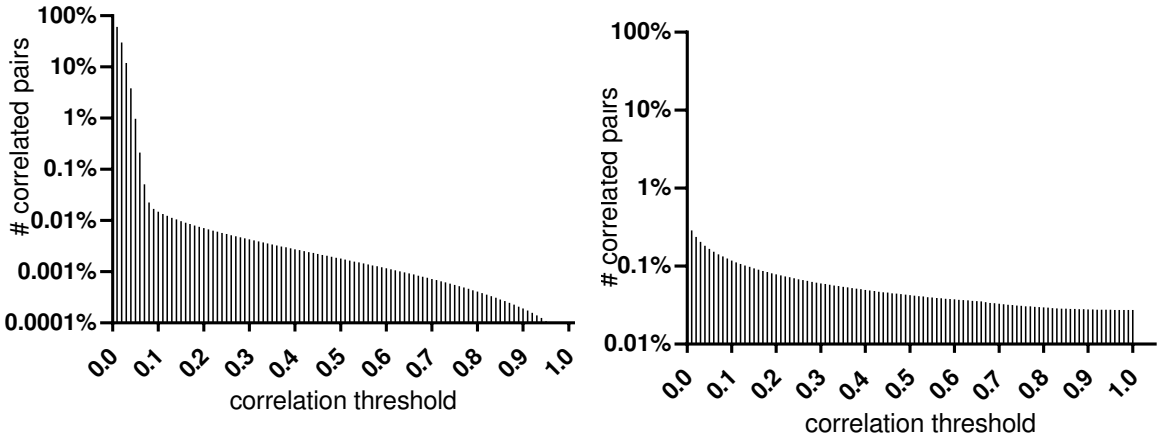
square of dimensionality in size) can become computationally very intensive. Notably, a dataset with *millions* of features would translate to *trillions* of correlations to be computed. Although some works have been proposed on fast correlation findings [164–166], it is still worth noting that such degree of extreme computational complexity poses a challenge that has received much less attention in the field of machine learning and data mining research.

To reveal and illustrate the complexity of such a challenge, the efforts to compute the correlations of two commonly used and well established datasets are analyzed in what follows, including `psoriasis` with 529,651 SNPs and `news20.binary` with 1,355,191 word frequencies[1]. Theoretically, it can be asserted that for a simple brute force approach, a total number of $\binom{m}{2}$ computations would be necessary to obtain the pairwise correlations between all features in the datasets considered, wherein m denotes the number of features. In particular, **0.14 and 0.92 trillion** correlation computations[2] are necessary on these datasets. On the `psoriasis` dataset, which has only 529,651 features, it already took us 20.6 days of wall clock time to compute the full pairwise correlations of the feature sets in LIBSVM format, on an Intel® Core™ i7-930 Processor. This clearly poses a serious impediment to the successful use of the feature correlation criterion on big dimensional datasets. Thus, there is a need for fresh computational and statistical learning paradigms to address such emerging challenge explicitly.

Fortunately, the detailed analyses on the well established datasets (which exhibit characteristics of big dimensionality) revealed that an extremely small portion (i.e., less than 0.1% for these two datasets considered) of the feature pairs have been found to be highly correlated. To illustrate this observation, the distributions of correlated feature pairs for the `psoriasis` and `news20.binary` datasets have been summarized in Figure 4.1(a) and Figure 4.1(b), respectively. In the figure, each bar denotes the percentage of feature pairs (the y-axis) that satisfies a given correlation threshold (as indicated on the x-axis). From the figure, the percentage of feature pairs is noted to decrease exponentially for increasing correlation threshold values. To be precise, 99.985% of the feature pairs in `psoriasis` and 99.882% in the `news20.binary` dataset have correlation values lower than

[1]Note that, in the real-world experimental studies, besides these two, datasets with up to 30 million dimensions are considered.

[2]To be exact, 140,264,826,075 and 918,270,645,645 correlation computations, respectively.



4.1.a: `psoriasis` 4.1.b: `news20.binary`

Figure 4.1: Distributions of correlated feature pairs in some established datasets, wherein each bar denotes the percentage of feature pairs (the y-axis) that has satisfied the given correlation threshold (as indicated on the x-axis), *i.e.*, $1 - \text{CDF}_i$. Note that the y-axis is in **log scale**.

the threshold of 0.1. This implies that majority of the feature pairs are uncorrelated or the features are sparsely correlated. In this thesis, this phenomenon is termed as "sparse correlation" and the aspiration is to exploit this sparse correlation that is made available through the "Blessings of Big Dimensionality" [59].

In this chapter, a novel learning approach is introduced to exploit the presence of sparse correlations for the efficient identifications of informative feature groups from datasets, especially in big dimensionality which involves general classification tasks. The proposed approach is a general feature grouping and selection framework, which considers an explicit incorporation of correlation measures as constraints in the learning model for different types of learning problem. An efficient embedded feature selection strategy, designed to remove large numbers of non-contributing correlations that could confuse the classifier, while identifying the informative feature groups, is then introduced. Extensive empirical studies on both synthetic and several real-world datasets comprising up to 30 million dimensions are subsequently conducted to assess and showcase the efficacy of our proposed approach. The core technical contributions of this chapter are summarized as

follows:

(i) The current work represents a first attempt to incorporate linear and non-linear feature correlation measures for feature grouping and selection in binary and one-class machine learning settings. The inclusion of correlation constraints among features in the learning model facilitates possible identifications of informative feature groups.

(ii) To achieve the goal, the notions of **support feature** and **affiliated feature** are explicitly defined. The former denotes the highly informative features with lower peer correlation, while the latter are features that are highly correlated to the support feature. Support-Affiliated Feature Groups are then established by an aggregation of the affiliated features that correspond to each support feature.

(iii) To identify the support-affiliated feature groups, in the proposed linear correlation setting, a generalized relation between the absolute pairwise Pearson's correlation coefficient and the discriminative score of features is derived, see Proposition 1. With such relationship established, the eliminations of uncorrelated feature pairs can thus be carried out without the need for a full correlation computation. It is worth noting that, this translates to a reduction in complexity on correlation computations, from $O(nm^2)$ to $O(m \log m + \mathcal{K}_a mn)$, where m is the dimensionality, n is the size of data, $\mathcal{K}_a$ is the total number of support features identified and $\mathcal{K}_a \ll \min(m, n)$ generally holds.

The remainder of this chapter is organized as follows. In Section 4.2, some of the core definitions used in this work are presented. Further, Section 4.3 introduces the proposed methodology to identify the support-affiliated feature groups effectively and efficiently. Then, the experimental setup and obtained results are presented in Section 4.4. The conclusive remark of the proposed framework is given in Section 4.5.

4.2 Preliminaries and Motivations

In this section, the core definitions and concepts that are used are presented throughout the rest of the chapter.

4.2.1 Feature Correlation Measures, *corr*$(\cdot,\cdot)$

In this subsection, both the linear and nonlinear instantiations of $corr(\cdot,\cdot)$ are illustrated.

Amongst various correlation measures, Pearson's correlation coefficient (PCC) is one of the most commonly used **linear** correlation measure [167]. The PCC for a pair of features $\mathbf{f}_j$ and $\mathbf{f}_k$, $\rho(\mathbf{f}_j, \mathbf{f}_k)$, can be defined as follows,

$$corr_{\mathbf{linear}}(\mathbf{f}_j, \mathbf{f}_k) \colon \rho(\mathbf{f}_j, \mathbf{f}_k) = \frac{cov(\mathbf{f}_j, \mathbf{f}_k)}{\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k}} \tag{4.1}$$

$$= \frac{(\mathbf{f}_j - \mu_{\mathbf{f}_j}\mathbf{1}')(\mathbf{f}_k - \mu_{\mathbf{f}_k}\mathbf{1}')'}{n\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k}}, \tag{4.2}$$

wherein $cov(\mathbf{f}_j, \mathbf{f}_k)$ designates the covariance of the two features. However, as its polarity does not affect the informativeness of a selected feature, the coefficient is hereinafter referred to the absolute form in the present study.

From linear to **nonlinear** correlations, Mutual Information (MI) represents a well established measure for feature selection [168], which takes the form of

$$I(\mathbf{f}_j; \mathbf{f}_k) = H(\mathbf{f}_j) + H(\mathbf{f}_k) - H(\mathbf{f}_j, \mathbf{f}_k) \tag{4.3}$$

(with $H(\cdot)$ denoting the entropy [169]) that measures the level of information sharing between feature $\mathbf{f}_j$ and $\mathbf{f}_k$ (i.e., $H(\mathbf{f}_j) \cap H(\mathbf{f}_k)$, where $H(\mathbf{f}) = -\sum_i p(f_i)\log_2 p(f_i)$). In feature selection, MI is typically used for assessing the ranking of the features in classification problem, i.e., a higher $I(\mathbf{f}; \mathbf{y})$ implies a higher devotion of feature $\mathbf{f}$ to class y.

MI is ranged as $[0, \infty]$ such that it is not a good quantization for pairwise correlation. Alternatively, the Symmetrical Uncertainty (SU) [142], which is a form of normalized MI has often been considered, which is defined by,

$$corr_{\mathbf{nonlinear}}(\mathbf{f}_j, \mathbf{f}_k) \colon U(\mathbf{f}_j, \mathbf{f}_k) = \frac{2I(\mathbf{f}_j; \mathbf{f}_k)}{H(\mathbf{f}_j) + H(\mathbf{f}_k)}. \tag{4.4}$$

Note that, both absolute PCC ($|\rho(\cdot,\cdot)|$) and SU ($U(\cdot,\cdot)$) are symmetrical measures that lie in $[0, 1]$. Without loss of generality, other forms of correlation measure with a range of $[0, 1]$ may also apply in $corr(\cdot,\cdot)$. Further, a high (low) value of $corr(\cdot,\cdot)$ indicates that the pair of features considered are strongly (weakly) correlated. Hence if two features are fully independent, their correlation shall be 0. On the other hand, when they are completely correlated to each other, namely, one feature can exactly predict the other, 1 follows.

4.2.2 Support and Affiliated Features

The core objective of traditional feature selection approaches is to identify a reduced feature subset of informative features [74, 162]. In contrast to previous studies, this chapter focuses on discovering the underlying interactions among informative features and capturing the salient characteristics within the data, based on the conception of sparse correlations and feature groupings, since such groupings can be useful to assist users in their interpretations of the data for further analysis. More specifically, the aim is at identifying feature groups through pairwise feature correlation among informative features. Though several prior works [76, 127] have highlighted the benefits of identifying feature groups (e.g., many biological studies have suggested that SNPs usually work in groups for some genetic activities), how to define the feature groups for general learning tasks is non-trivial.

In the present study, the interest is to identify a sparse feature subset of support features, while discovering the feature groups. Each feature group comprises a parent support feature with affiliated features as children that are strongly correlated to it. In what follows, the definitions of the support feature and affiliated feature are given, which form the basis of the current work.

Definition 1. *A Support Feature (SF_k) denotes the most informative (discriminative) feature w.r.t. the output labels among the residual features. All of the support features identified, as depicted in full circles of Figure 4.2, are uncorrelated or weakly correlated to one another. In other words, the pairwise feature correlation between a pair of support features is lower than the predefined precision threshold ε (i.e., $corr(SF_j, SF_k) < \varepsilon$).*

Definition 2. *An Affiliated Feature (AF_{k_s}) should also be an informative feature, which shares similar predictive capability with the associated support feature SF_k, and is strongly correlated with SF_k (i.e., $corr(SF_k, AF_{k_s}) \geq \varepsilon$). The dotted circles of Figure 4.2 showcase this type of features.*

With the above definitions, the interest of our current work is to discover support-affiliated feature groups that takes the form of Figure 4.2 from various datasets in different machine learning settings.

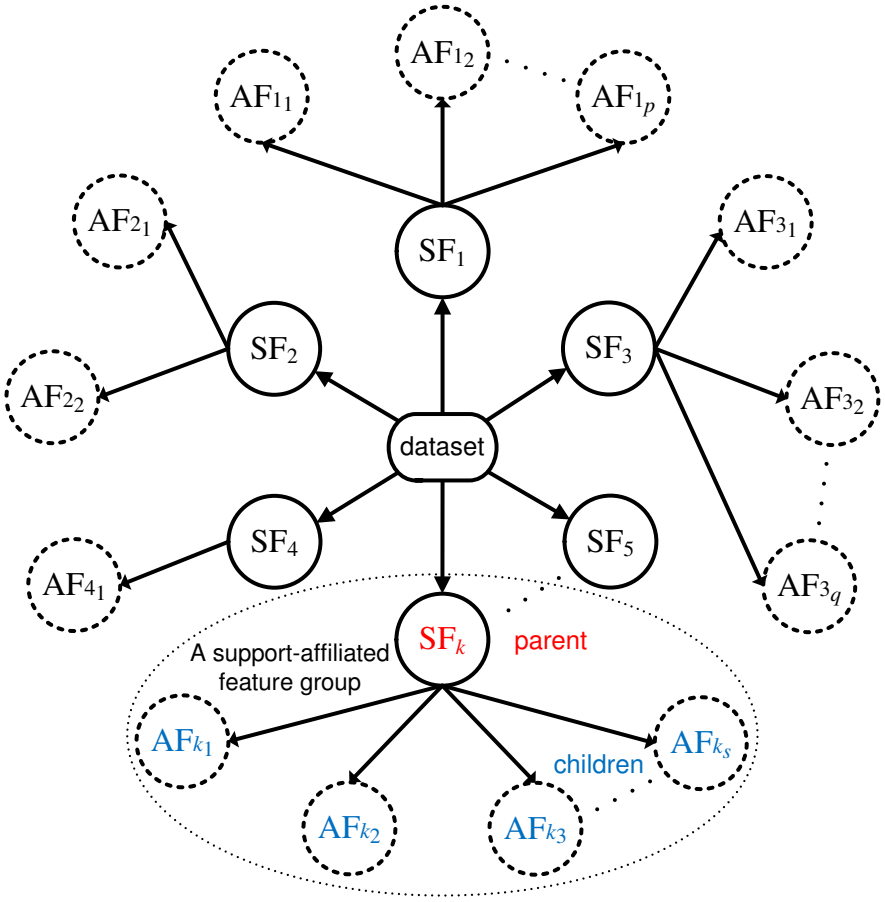


Figure 4.2: Structural relationship of support-affiliated feature groups (denoted using dotted ellipse). SF_k: support feature (parent denoted using full circle), $\mathrm{AF}_{k_{1,2,3,\ldots,s}}$: affiliated features (children of SF_k as denoted by dotted circles).

4.3 Group Discovery Machine

In this section, a novel feature grouping and selection method, labeled here as the **Group Discovery Machine** (GDM), is introduced for the discovery of support-affiliated feature groups in various machine learning tasks. The essential backbone of the GDM is a sparse SVM with an efficient Quadratically Constrained Quadratic Programming (QCQP) solver. An explicit incorporation of the pairwise linear/nonlinear correlation measures is introduced as constraints in the learning model to discover the appropriate support-affiliated feature groups.

4.3.1 General Correlation Constraints

Similar to the idea of feature indicator in [79], a vector $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_m]' \in \{0,1\}^m$ is introduced to define whether a corresponding SF is selected ($\delta_j = 1$) or not ($\delta_j = 0$), such that the decision function is given by: $f(\mathbf{x}) = \mathbf{w}'(\mathbf{x} \odot \boldsymbol{\delta})$, where the vector $\mathbf{w} \in \mathbb{R}^m$ denotes weight vector. To limit the number of selected features to be lower than $\mathcal{K}_a$, the ℓ_0-constraint $\|\boldsymbol{\delta}\|_0 \leq \mathcal{K}_a$ is imposed for the purpose of feature selection. Further, to constrain the correlation among the selected features, the following constraint on $\boldsymbol{\delta}$ is explicitly introduced here as

$$\delta_j \delta_k = 0 \text{ if } |corr(\mathbf{f}_j, \mathbf{f}_k)| \geq 1 - \tau, \forall j, k \text{ with } j \neq k. \tag{4.5}$$

With this constraint, the feature pair is regarded as uncorrelated if their correlation coefficient falls below the bound $(1-\tau)$, where $\tau \in [0, 0.5]$.[3] Next, $\boldsymbol{\Delta} = \{\boldsymbol{\delta} | \sum_{j=1}^{m} \delta_j \leq \mathcal{K}_a;\ \delta_j \in \{0,1\};\ \delta_j \delta_k = 0 \text{ if } |corr(\mathbf{f}_j, \mathbf{f}_k)| \geq 1 - \tau, \forall j, k \text{ with } j \neq k\}$ is defined as the domain for $\boldsymbol{\delta}$. Further, (4.5) explicitly defines $\binom{m}{2} \Rightarrow O(m^2)$ quadratic constraints with m numbers of integer variables. As previously noted, our present task is inclined to solve problems with *millions of dimensions*, which translates to *trillion quadratic constraints*. Moreover, seeking the solution $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ involves a process of combinatorial subset selection, resulting in extremely high computational cost, especially on big dimensional data. In what follows, the proposed approach is described in detail to deal with the trillion correlation constraints that arise.

4.3.2 Proposed Formulation

In GDM, the interest is to find a large margin decision function $f(\mathbf{x})$ for robust prediction, and seamlessly identify the informative yet uncorrelated feature subset that satisfies the constraints defined in (4.5). For the purpose of simplicity, the square hinge loss in SVM is considered, thus arriving at the following optimization problems:

$$\begin{aligned} \min_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \min_{\mathbf{w}, \gamma, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 - \gamma + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2 \\ s.t. \quad & y_i \mathbf{w}'(\mathbf{x}_i \odot \boldsymbol{\delta}) \geq \gamma - \xi_i \quad i = 1, \cdots, n, \end{aligned} \tag{4.6}$$

[3] If pairwise feature correlation is below the mean of the interval (i.e., 0.5 of $corr(\cdot, \cdot)$), it should not be considered as high correlated in most of statistic methods.

where $\xi_i \geq 0$ is the slack variable, $\gamma/\|\mathbf{w}\|$ denotes the margin and C is a tradeoff parameter to regulate the function complexity $\|\mathbf{w}\|_2^2$ and the training error (ξ_i's). Note, as discussed earlier, the optimization problem in (4.6) with constraints defined in (4.5) is a challenging problem, as a result of the explosion in the number of constraints involving big dimensional data.

4.3.3 Solving the Problem Iteratively with Cutting Plane Algorithm

Cutting planes are a major component of the mixed integer linear optimization solver for accelerating the progress by removing fractional solutions. Recently, the *Cutting Plane Algorithm* has reported much success in many problems involving vast varieties of constraints, including SVM training [170], structure prediction [170], maximum margin clustering [171] and so on.

Taking the cue, here the problem (4.6) is solved by incorporating a cutting plane approach. To begin, the inner minimization section of problem (4.6) considers a dual form of SVM w.r.t. $\mathbf{w}, \gamma$ and ξ_i. Thus (4.6) becomes a minimax saddle-point problem. Inspired by applying the minimax optimization theory, a tight convex relaxation to problem (4.6) can be attained, which takes the form of a Quadratically Constrained Quadratic Programming (QCQP) problem:

$$\min_{\boldsymbol{\alpha}\in\mathcal{A},\theta} \theta : \theta \geq g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}), \forall \boldsymbol{\delta} \in \boldsymbol{\Delta} \quad \text{or} \quad \min_{\boldsymbol{\alpha}\in\mathcal{A}} \max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \tag{4.7}$$
$$\text{define} \;\; g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) = \frac{1}{2}\Big\| \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta}) \Big\|^2 + \frac{1}{2C}\boldsymbol{\alpha}'\boldsymbol{\alpha},$$

wherein $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]'$ is the vector of dual variables, $\mathcal{A} = \{\boldsymbol{\alpha} \big| \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0, \forall i = 1, \cdots, n\}$ defines the domain of $\boldsymbol{\alpha}$, and θ is the upper bound of $g_{\boldsymbol{\delta}}(\cdot)$. Nevertheless, since there are as many as $(\sum_{i=0}^{\mathcal{K}_a} \binom{m}{i})$ quadratic constraints in problem (4.7), the problem remains to be plagued with computational complexity issues. The *Cutting-Set methods* developed in [155] considers a general worst-case convex optimization problem with arbitrary dependence on the uncertain parameters. Hence, rather than solving the original problem which involves a vast number of constraints, cutting-set is used here to generate a subset of active constraints in an iterative manner. This leads to a relaxed

optimization problem with the current constraint set considered. At the t^{th} iteration, $\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \geq g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha}), \forall \boldsymbol{\delta}^t \in \boldsymbol{\Delta}$ holds, and correspondingly $\boldsymbol{\delta}^t$ is constrained by a $\mathcal{K}_b$ (i.e., support feature size per iteration), where $\sum_t \mathcal{K}_b^t = \mathcal{K}_a$ generally holds. Thus, for a reduced active constraint set $\Lambda \subset \boldsymbol{\Delta}$, the lower bound approximation of (4.7) can be obtained as $\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \geq \max_{t=1,\ldots,T} g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha})$ with $T = |\Lambda|$, where T is the maximum number of constraints (iterations) imposed. This leads to solving a reduced problem of (4.7) that takes the form

$$\min_{\boldsymbol{\alpha}\in\mathcal{A},\theta} \theta \ : \ \theta \geq g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha}), \ \ \forall \ \boldsymbol{\delta}^t \in \Lambda. \tag{4.8}$$

4.3.4 Training with Multiple Kernel Learning

In the research field of kernel methods, several efficient and elegantly designed *Multiple Kernel Learning* (MKL) approaches have been proposed over the recent years. For instance, Lanckriet *et al.* first proposed the use of Quadratically Constrained Quadratic Programming (QCQP) in MKL [172]. Recently, Sonnenburg *et al.* proposed a semi-infinite linear programming formulation, which enables MKL to be iteratively solved with standard SVM solver as well as linear programming [173]. Rakotomamonjy *et al.* also proposed a related SimpleMKL algorithm using the reduced gradient descent procedure [174]. In this subsection, MKL optimization technique is considered for solving the problem defined in (4.8), wherein the aim is to jointly learn both the kernel and SVM parameters, or briefly, to identify the most appropriate kernel for addressing the task on hand [174].

Since problem (4.8) follows a convex QCQP problem, μ^t is introduced as the dual variable of each constraint. The Lagrangian function then takes the form of

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}) = -\theta + \sum_{t,\boldsymbol{\delta}^t\in\Lambda} \mu^t(\theta - g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha})). \tag{4.9}$$

Setting the derivative w.r.t. θ as zero, $\sum \mu^t = 1$ can be attained. $\boldsymbol{\mu}$ is set as the vector of μ^t's, and $\mathcal{U} = \{\boldsymbol{\mu} | \sum \mu^t = 1, \mu^t \geq 0\}$ defines the domain of $\boldsymbol{\mu}$. Consequently, the Lagrangian function $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu})$ can be rewritten as,

$$\begin{aligned} & \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{U}} \sum_{\boldsymbol{\delta}^t\in\Lambda} -\mu^t g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha}) \qquad (4.10) \\ = \quad & \min_{\boldsymbol{\mu}\in\mathcal{U}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}(\boldsymbol{\alpha}\odot\mathbf{y})'(\sum_{\boldsymbol{\delta}^t\in\Lambda} \mu^t \mathbf{X}_t\mathbf{X}_t' + \frac{1}{C}\mathbf{I})(\boldsymbol{\alpha}\odot\mathbf{y}), \end{aligned}$$

Input: Dataset $\mathcal{D}(\mathbf{X}, \mathbf{y})$, zero-one vector $\boldsymbol{\delta} \in \mathbb{R}^m$, support feature size per iteration $\mathcal{K}_b$ and correlation threshold τ.
Output: Index set $\mathcal{SF}$ for SFs and $\mathcal{AF}$ for AFs.
Initialization: $\boldsymbol{\alpha} = \mathbf{1}/n, \boldsymbol{\delta} = \mathbf{0}^m, \mathcal{S} = \varnothing$ and $\mathcal{Q} = \varnothing$.
for $t = 1$ **to** T **do**
1: Call $\boldsymbol{\delta}^t = \text{CRM}(\mathcal{D}, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}, \mathcal{AF})$
2: Set $\Lambda = \Lambda \cup \{\boldsymbol{\delta}^t\}$ and solve (4.8), while updating $\boldsymbol{\alpha}^{t+1}$
3: Quit if the objective value is convergent.
end for

Algorithm 2: Group Discovery Machine - GDM($\mathbf{w}, \boldsymbol{\delta}, \mathcal{D}$)

where $\mathbf{X}_t = [\mathbf{x}_1 \odot \boldsymbol{\delta}^t, \cdots, \mathbf{x}_n \odot \boldsymbol{\delta}^t]'$, and the equation follows on account of the fact that the objective function is concave in $\boldsymbol{\alpha}$ and convex in $\boldsymbol{\mu}$. The recently developed MKL is ideal for solving the resultant minimax problem (4.10) [172, 174], where the kernel matrix $\sum_{\boldsymbol{\delta}^t \in \Lambda} \mu^t \mathbf{X}_t \mathbf{X}_t'$ to be learned is a convex combination comprising $|\Lambda|$ number of base kernel matrices ($\mathbf{X}_t \mathbf{X}_t'$), each of which is constructed from a feasible $\boldsymbol{\delta}^t \in \Lambda$.

To summarize, the steps for solving the proposed problem are outlined in Algorithm 2, wherein some of the notations are explained thereinafter. Specifically, for each iteration of Algorithm 2, one needs to figure out the worst case analysis (i.e., finding the most violated constraint $\boldsymbol{\delta}^t$) of Problem (4.7) [155, 175], which is described in the following subsections 4.3.5 - 4.3.7. The obtained $\boldsymbol{\delta}^t$ is then appended into the active constraint set Λ, which forms a subset of $\boldsymbol{\Delta}$. Last but not least, the problem w.r.t. Λ can be solved via efficient QCQP solvers [163].

4.3.5 Correlation Redundancy Matching (CRM): finding the most violated constraints $\boldsymbol{\delta}$

In this subsection, the worst case analysis of problem (4.7), which plays a key role in *Cutting Plane Algorithm* [155] is presented. In the current problem setting, problem (4.8) is transformed into solving the following integer optimization problems:

$$\max_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \left\| \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta}) \right\|^2 \tag{4.11}$$

In general, solving such a problem is considered NP-hard. However, since one can obtain $\| \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta}) \|^2 = \| \sum_{i=1}^n (\alpha_i y_i \mathbf{x}_i) \odot \boldsymbol{\delta} \|^2 = \sum_{j=1}^m s_j^2 \delta_j$, where s_j is defined as

the *feature discriminative score* that follows

$$s_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij} = \sum_{i=1}^{n} \alpha_i y_i f_{ji} = \mathbf{f}_j \tilde{\boldsymbol{\alpha}} \tag{4.12}$$

with $\tilde{\boldsymbol{\alpha}} = [\alpha_1 y_1, \ldots, \alpha_n y_n]'$, indicating that the informative features should accord with features of largest absolute value feature score $|s_j|$'s. Moreover, recall that the correlation measures are embedded in $\boldsymbol{\delta}$, thus a natural question arises: considering all the correlated features, which one poses higher importance to the output labels?

To address this question, first of all, it is necessary to offer the instantiations of SF and AF in the proposed GDM. As discussed previously, SFs refer to the most informative features with relatively low pairwise correlations in this work. AFs, on the other hand, refer to the correlated features associated with each SF correspondingly. The parent-child structured relationship between SFs and AFs is illustrated in Figure 4.2.

Definition 3. *SF and AF in GDM: Given any exemplar vector* $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^n$ *and a collection of feature vectors* $\{\mathbf{f}_i\}$*, where* $\mathbf{f}_i' \in \mathbb{R}^n$*. The SF is given by* $\max_i |\mathbf{f}_i \tilde{\boldsymbol{\alpha}}|$ *for the given* $\tilde{\boldsymbol{\alpha}}$*. The remaining correlated features in* $\{\mathbf{f}_j\}$ *w.r.t.* $\mathbf{f}_i$ *(with* $|corr(\mathbf{f}_i, \mathbf{f}_j)| \geq 1 - \tau$*) then denote the AFs.*

For the sake of conciseness, let $\mathcal{SF}$ be the index set of the SFs and here a data structure $\mathcal{AF} = \{\mathcal{G}_j\}$ is introduced to represent the hierarchical structure of features, where $\mathcal{G}_j$ denotes the index set of the AFs for the j^{th} SF. In this manner, all the correlated features can be identified and archived instead of omitting them.[4] Moreover, based on the definitions above, once a support feature (SF_j) is identified (i.e., the feature with the largest $|s_j|$), all relevant features (AF_{j_s}) that correlate with SF_j then become the affiliated features that correspond to it. As the present proposed method discovers the correlated feature groups, it is labelled as the **Group Discovery Machine** (GDM) here. Note that, alternatively, one could employ a brute-force approach to search across all features and pairwise correlations to identify all feature groups that achieves the similar goal. However, such a scheme (i.e., mRMR [161]) can be computationally intensive even with small dataset and would become computational intractable on big dimensional data. For

[4]The practice of existing works in the literature is to omit all correlated features, *i.e.*, redundancy reduction.

the details on the intensiveness of a brute-force approach, please refer to the Section i of the Appendix.

Seeking for the most violated constraints $\boldsymbol{\delta}$ is then termed here as *Correlation Redundancy Matching* (CRM) procedure. In CRM, once an SF is identified, the AFs are isolated from the rest of the features based on their correlations w.r.t. the SF. The above procedure is repeated until a maximum of $\mathcal{K}_b$ unique support features are identified in each iteration.

4.3.6 CRM with Linear Correlation $corr_{\textbf{linear}}(\cdot,\cdot)$

In this section, the way of feature grouping with linear correlation is illustrated, i.e., using PCC $\rho(\cdot,\cdot)$ as the correlation measure. To this end, firstly, a proposition is presented to prove the case of linear correlation, which serves as a generalization of that previously presented in [163] on assumption made pertaining to data normalization: for a group of strongly correlated features, if one of them is informative to the output labels, all of them can be treated identically (i.e., all of them will make positive contributions to the output label).

Proposition 1. *Given a nonzero column vector* $\tilde{\boldsymbol{\alpha}}$ *and any two feature vectors* $\mathbf{f}_1$ *and* $\mathbf{f}_2$*, suppose their absolute PCC* $|corr(\mathbf{f}_1,\mathbf{f}_2)| = |\rho(\mathbf{f}_1,\mathbf{f}_2)| \geq (1-\tau)$*, then* $|\,|\mathbf{f}_1\tilde{\boldsymbol{\alpha}}| - |\mathbf{f}_2\tilde{\boldsymbol{\alpha}}|\,| \leq \sqrt{n(\Delta_\sigma^2+\Delta_\mu^2+2\tau\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2})}\|\tilde{\boldsymbol{\alpha}}\|$ *holds, where the correlation parameter* $\tau \in (0,1)$ *and* $\Delta_\sigma = \sigma_{\mathbf{f}_1} - \sigma_{\mathbf{f}_2}$ *while* $\Delta_\mu = \mu_{\mathbf{f}_1} - \mu_{\mathbf{f}_2}$*.*

Proof: Firstly, the computation of the absolute correlation is arrived for any feature pair, *i.e.*, $|corr(\mathbf{f}_1,\mathbf{f}_2)|$. Using PCC, as what Equation (4.1) shows, one can arrive at

$$\begin{aligned}|\rho(\mathbf{f}_1,\mathbf{f}_2)| &= |\frac{(\mathbf{f}_1-\mu_{\mathbf{f}_1}\mathbf{1}')(\mathbf{f}_2-\mu_{\mathbf{f}_2}\mathbf{1}')'}{n\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2}}| \\ &= |\frac{\mathbf{f}_1\mathbf{f}_2' - \mu_{\mathbf{f}_1}\mathbf{1}'\mathbf{f}_2' - \mu_{\mathbf{f}_2}\mathbf{1}\mathbf{f}_1 + n\mu_{\mathbf{f}_1}\mu_{\mathbf{f}_2}}{n\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2}}|.\end{aligned}$$

Assuming that $|\rho(\mathbf{f}_1,\mathbf{f}_2)| \geq 1-\tau$, the following inequality holds,

$$|\rho(\mathbf{f}_1,\mathbf{f}_2)| = |\frac{\mathbf{f}_1\mathbf{f}_2' - n\mu_{\mathbf{f}_1}\mu_{\mathbf{f}_2}}{n\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2}}| \geq 1-\tau.$$

As a common sense that inequality with absolute value $|\cdot|$ has two sides. Here the proof on the positive side is showcased (*i.e.*, $\mathbf{f}_1$ and $\mathbf{f}_2$ are positive correlated), and vise versa.

Based on this, since $n\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2}$ is always non-negative and suppose $\sigma_{\mathbf{f}_i} > 0$ holds for all the features,

$$\mathbf{f}_1\mathbf{f}_2' \geq n[(1-\tau)\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2} + \mu_{\mathbf{f}_1}\mu_{\mathbf{f}_2}]$$

can be attained. From above, with $\|\mathbf{f}_j\|^2 = n(\sigma_{\mathbf{f}_j}^2 + \mu_{\mathbf{f}_j}^2)$, one arrives at

$$\begin{aligned} & \|\mathbf{f}_1 - \mathbf{f}_2\|^2 = \|\mathbf{f}_1\|^2 + \|\mathbf{f}_2\|^2 - 2\mathbf{f}_1\mathbf{f}_2' \\ = \quad & n(\sigma_{\mathbf{f}_1}^2 + \sigma_{\mathbf{f}_2}^2 + \mu_{\mathbf{f}_1}^2 + \mu_{\mathbf{f}_2}^2) - 2\mathbf{f}_1\mathbf{f}_2' \\ \leq \quad & n[(\sigma_{\mathbf{f}_1} - \sigma_{\mathbf{f}_2})^2 + (\mu_{\mathbf{f}_1} - \mu_{\mathbf{f}_2})^2 + 2\tau\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2}] \\ = \quad & n(\Delta_\sigma^2 + \Delta_\mu^2 + 2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k}) \end{aligned}$$

based on the definitions of $\Delta_\sigma = \sigma_{\mathbf{f}_1} - \sigma_{\mathbf{f}_2}$ and $\Delta_\mu = \mu_{\mathbf{f}_1} - \mu_{\mathbf{f}_2}$, correspondingly, with Cauchy-Schwarz inequality, the inequality below holds,

$$\begin{aligned} & |\,|\mathbf{f}_1\tilde{\boldsymbol{\alpha}}| - |\mathbf{f}_2\tilde{\boldsymbol{\alpha}}|\,| \\ \leq \quad & |\mathbf{f}_1\tilde{\boldsymbol{\alpha}} - \mathbf{f}_2\tilde{\boldsymbol{\alpha}}| = |(\mathbf{f}_1 - \mathbf{f}_2)\tilde{\boldsymbol{\alpha}}| \leq \|\mathbf{f}_1 - \mathbf{f}_2\|\|\tilde{\boldsymbol{\alpha}}\| \\ \leq \quad & \sqrt{n(\Delta_\sigma^2 + \Delta_\mu^2 + 2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|. \end{aligned}$$

On the other hand, in the case of negative correlation, a positive correlated vector $\hat{\mathbf{f}}_2 = -\mathbf{f}_2$ is defined, hence the proof will follow a similar form with the derivation above. This completes the proof.

The above results state that if two feature vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ are highly correlated, their distance measure (or correlation) to any exemplar vector $\tilde{\boldsymbol{\alpha}}'$ will be close to one another. In what follows, a theorem is presented to illustrate that in practice one can address the linear pairwise feature correlation by scanning only a small subset of the features on big dimensional problems.

Theorem 1. *Given a nonzero column vector $\tilde{\boldsymbol{\alpha}}$ and any two feature vectors $\mathbf{f}_j$ and $\mathbf{f}_k$, suppose $|\rho(\mathbf{f}_j, \mathbf{f}_k)| \geq (1-\tau)$ and $\mathbf{f}_j$ is the support feature (i.e., $\mathbf{f}_k$ is qualified as an affiliated feature of $\mathbf{f}_j$) with feature score $|s_j| = |\mathbf{f}_j\tilde{\boldsymbol{\alpha}}|$ based on Equation (4.12), then the feature score of $\mathbf{f}_k$ satisfies $|s_k| \geq |s_j| - \sqrt{n(\Delta_\sigma^2 + \Delta_\mu^2 + 2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|$ with $\Delta_\sigma = \sigma_{\mathbf{f}_1} - \sigma_{\mathbf{f}_2}$ and $\Delta_\mu = \mu_{\mathbf{f}_1} - \mu_{\mathbf{f}_2}$.*

1: Initialize $k = 1$ and denote $\varrho\|\tilde{\boldsymbol{\alpha}}\|$ as the bound arrived from Theorem 1. Set the output $\boldsymbol{\delta}^t = \mathbf{0}$.
2: Compute feature score vector $\mathbf{s}$ according to (4.12) and sort $|s_j|$ in descending order, record the feature ranking list as $\mathcal{E}$.
while $\|\boldsymbol{\delta}^t\|_0 < \mathcal{K}_b$ **do**
  Pick the k^{th} feature $\mathbf{f}_z$ from $\mathcal{D}$, where $z = \mathcal{E}(k)$
  if $(\mid |s_z|^t - |s_j|^t \mid > \varrho_{z,j}\|\tilde{\boldsymbol{\alpha}}\|^t$ with all existed SF $\mathbf{f}_j)$ **then**
    $\mathcal{SF} = \mathcal{SF} \cup \{z\}$ and $\delta_z^t = 1$ ($\mathbf{f}_z$ is set as new SF)
  else
    For the SFs that satisfy $(\mid |s_z|^t - |s_j|^t \mid \leq \varrho_{z,j}\|\tilde{\boldsymbol{\alpha}}\|^t)$, compute $\rho(\mathbf{f}_z, \mathbf{f}_j)$.
    if $(\exists j, \rho(\mathbf{f}_z, \mathbf{f}_j) \geq 1 - \tau)$ **then**
      $\mathcal{AF}.\mathcal{G}_j = \mathcal{AF}.\mathcal{G}_j \cup \{z\}$ ($\mathbf{f}_z$ is set as new AF for SF_j)
    else
      $\mathcal{SF} = \mathcal{SF} \cup \{z\}$ and $\delta_z^t = 1$ ($\mathbf{f}_z$ is set as new SF)
    end if
  end if
  Set $k = k + 1$
end while

Algorithm 3: CRM$(\mathcal{D}, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}, \mathcal{AF})$ with PCC

Proof: From Proposition 1, $\mid |\mathbf{f}_j\tilde{\boldsymbol{\alpha}}| - |\mathbf{f}_k\tilde{\boldsymbol{\alpha}}| \mid \leq \sqrt{n(\Delta_\sigma^2 + \Delta_\mu^2 + 2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|$ can be obtained under $|\rho(\mathbf{f}_j, \mathbf{f}_k)| \geq 1 - \tau$. Further, since $\mathbf{f}_j$ is the support feature, which has the highest feature score among all other features, so $|\mathbf{f}_j\tilde{\boldsymbol{\alpha}}| \geq |\mathbf{f}_k\tilde{\boldsymbol{\alpha}}|$. Correspondingly, $|\mathbf{f}_k\tilde{\boldsymbol{\alpha}}| = |s_k| \geq |s_j| - \sqrt{n(\Delta_\sigma^2 + \Delta_\mu^2 + 2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|$ holds. This completes the proof.

The above theorem states that if two features are strongly correlated, their scores will be close to one another. In other words, for a given support feature $\mathbf{f}_j$, all other features with scores that fall below the arrived bound at the correlation level of $(1 - \tau)$, shall not be considered as the affiliated features of $\mathbf{f}_j$. Correspondingly, this facilitates possible eliminations of vast numbers of uncorrelated features without the need to undergo extensive correlation computations. The details of seeking $\boldsymbol{\delta}$ using PCC is then illustrated in Algorithm 3, and GDM that employs the PCC is termed here as GDM-PCC.

4.3.7 CRM with Nonlinear Correlation $corr_{\mathbf{nonlinear}}(\cdot, \cdot)$

Besides linear correlation, nonlinear relationship is also considered as fundamental to many statistical, physical and biological phenomena [176]. Among the nonlinear correlation measures, the normalized mutual information is considered as important, even

1: Initialize $k = 1$ and set the output $\boldsymbol{\delta}^t = \mathbf{0}$.
2: Compute feature score vector $\mathbf{s}$ and sort $|s_j|$ in descending order, record the feature ranking list as $\mathcal{E}$.
while $\|\boldsymbol{\delta}^t\|_0 < \mathcal{K}_b$ **do**
  Pick the k^{th} feature $\mathbf{f}_z$ from $\mathcal{D}$, where $z = \mathcal{E}(k)$
  if $(\exists j, U(\mathbf{f}_z, \mathbf{f}_j) \geq 1 - \tau)$ for SF_j **then**
    $\mathcal{AF}.\mathcal{G}_j = \mathcal{AF}.\mathcal{G}_j \cup \{z\}$ ($\mathbf{f}_z$ is set as new AF for SF_j)
  else
    $\mathcal{SF} = \mathcal{SF} \cup \{z\}$ and $\delta_z^t = 1$ ($\mathbf{f}_z$ is set as new SF)
  end if
  Set $k = k + 1$
end while

Algorithm 4: CRM($\mathcal{D}, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}, \mathcal{AF}$) with SU

touted as "a correlation for the 21st century" [177]. Here, GDM is shown to offer flexibility and room for nonlinear correlation measure to handle more complex tasks. For the sake of brevity, in this section, the symmetrical uncertainty (SU) is considered as the normalized MI in GDM (i.e., $corr(\cdot) = U(\cdot)$) and the method is termed here as GDM-SU, correspondingly[5]. Algorithm 4 summarizes the pseudo code of feature grouping and selection with nonlinear correlation SU.

4.3.8 CRM with Specific Correlation Constraints

In this subsection, a biological case is used to showcase the flexibility of the proposed CRM. Notably, the specific correlation constraint has been designed using the prior knowledge in order to enhance the classification performance.

In human genome, gene is the basic unit representing the organism's hereditary information, and as is well-known, one DNA or RNA (Ribonucleic acid) molecule carries numerous genes. Usually tens of thousands of base pairs when grouped together then form a gene. And since SNP is a base pair mutation, it is more reasonable to consider the SNP correlation at the gene level rather than simply regarding the datasets as raw feature pools. As what Figure 4.3 illustrates, gene information can be useful for localizing the selected SFs and AFs, i.e., although high correlation between feature SNPs $\mathbf{f}_2$ and $\mathbf{f}_9$, they should be separated by the different gene that they belong to, while gene-localized

[5]Note that other forms of nonlinear correlation measure that satisfies the property also applies.

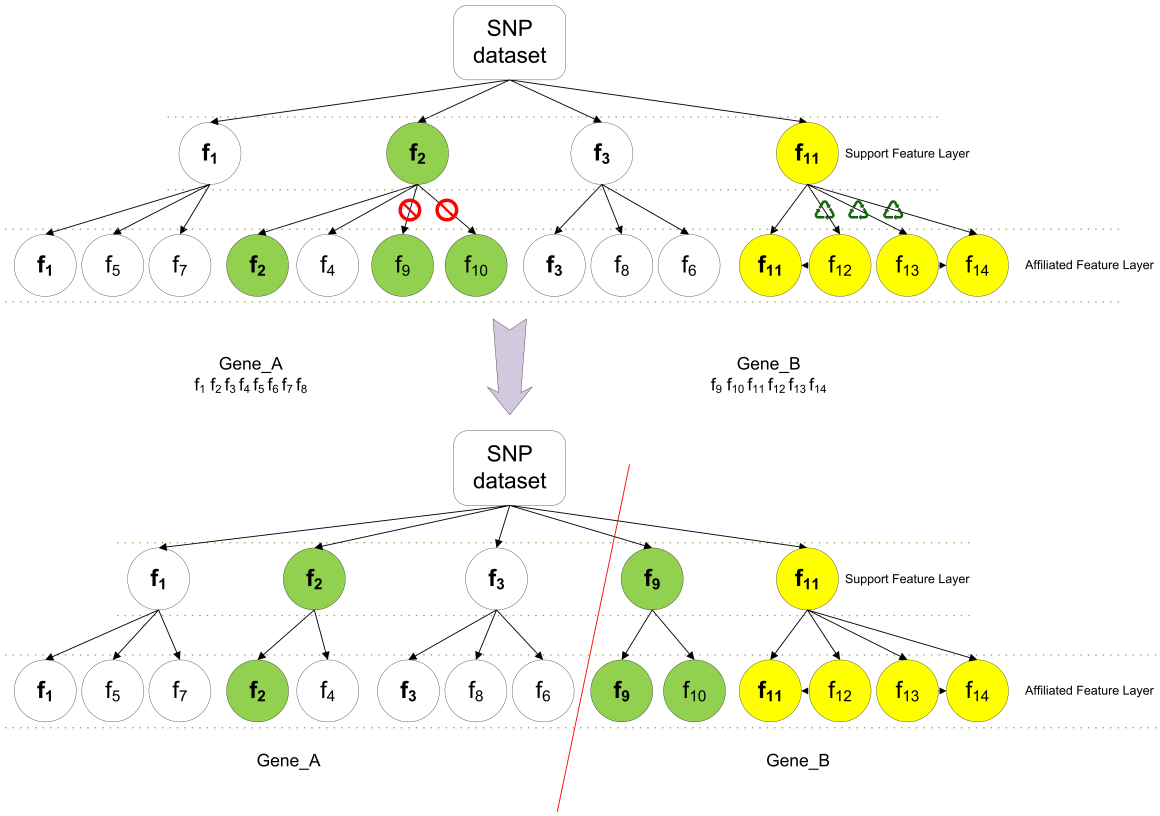


Figure 4.3: Maximizing the utilities of truly useful SNPs, based on gene-localized SNP correlation.

SNPs ($\mathbf{f}_{11}$, $\mathbf{f}_{12}$, $\mathbf{f}_{13}$ and $\mathbf{f}_{14}$) are maintained. However, in SNP study especially on the encoded dataset, it is often hard to differentiate the genes correctly, owing to the reason that only the base pairs of SNPs are available in the data. To account for this shortfall, a formulation that takes into considerations on the relative position of the SNPs within a gene and/or the molecule is proposed. In particular, the correlation of pairwise SNPs are weighted according to how far they are apart from each other, w.r.t. their positional distance on the molecule.

Here, an exponential multiplier is designed to augment the basic correlation criterion, which weights according to how far SNPs differ in centimorgan distance within a molecule, as

$$\delta_j \delta_k = 0, |\rho(\mathbf{f}_j, \mathbf{f}_k) e^{-\frac{|d_j - d_k|}{\beta}}| \geq 1 - \tau, \forall j \neq k, \tag{4.13}$$

where d_j and d_k are the SNP indices of feature j and k indicated from the dataset, respectively, and β is the penalization parameter. A smaller value of β will result in a

larger penalty and consequently, a lower confidence to match the correlation threshold of $1-\tau$.

Although there exists some highly correlated SNPs from different genes, this correlation may not be that significant, since gene is the basic unit carrying hereditary information as aforementioned. With the new constraint, the correlated SNPs with relatively close centimorgan distance is expected to be effective correlation (i.e., SNPs with high possibility to be in the same gene), while far apart correlated SNPs will be disregarded.

4.3.9 Convergency Analysis of GDM

In this subsection, the convergence analysis is conducted by introducing some Theorems and Propositions.

Theorem 2. *Let $(\boldsymbol{\alpha}^*, \theta^*)$ be the global optimal pair of (4.7), define*

$$\beta^k = \max_{1\leq i\leq k} g_{\boldsymbol{\delta}^i}(\boldsymbol{\alpha}^k) = \min_{\boldsymbol{\alpha}\in\mathcal{A}} \max_{1\leq i\leq k} g_{\boldsymbol{\delta}^i}(\boldsymbol{\alpha})$$

$$\textit{and} \quad \varphi^k = \min_{1\leq j\leq k} g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j),$$

where k denotes the number of iterations, then one can arrive at $\beta^k \leq \theta^* \leq \varphi^k$. And with an increasing k, β^k is monotonically increasing while φ^k is monotonically decreasing.

Proof: Firstly, $\theta^* = \min\limits_{\boldsymbol{\alpha}\in\mathcal{A}} \max\limits_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha})$. For a fixed feasible $\boldsymbol{\alpha}$, one can arrive at $\max\limits_{\boldsymbol{\delta}\in\mathcal{C}^k} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \leq \max\limits_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha})$, which has a equal form to $\min\limits_{\boldsymbol{\alpha}\in\mathcal{A}} \max\limits_{\boldsymbol{\delta}\in\mathcal{C}^k} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \leq \min\limits_{\boldsymbol{\alpha}\in\mathcal{A}} \max\limits_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha})$, *i.e.*, $\beta^k \leq \theta^*$. On the other hand, for $\forall j = 1, \cdots, k$, where j denotes the j^{th} iteration (with max number of k iterations), $g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j) = \max\limits_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}^j)$, thus $(\boldsymbol{\alpha}^j, g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j))$ is a feasible and possible solution pair for problem (4.7). Consequently, since $\theta^* \leq g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j)$ holds for $j = 1, \cdots, k$, the inequality $\theta^* \leq \varphi^k = \min\limits_{1\leq j\leq k} g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j)$ can be established. Obviously, with increasing number of iterations k, the corresponding subset $\mathcal{C}^k$ will be monotonically increasing as well, so is β^k, while φ^k is monotonically decreasing. This completes the proof.

Proposition 2. *With the proposed Correlation Redundancy Matching algorithm, the most violated constraint selection problem (4.11) can be solved with the most informative feature selected.*

Proof: From Algorithm 3, once the most informative $\mathbf{f}_z$ is identified as the first support feature, all corresponding correlated features of $\mathbf{f}_z$ are subsequently identified and archived in $\mathcal{AF}.\mathcal{G}_z$. This implies that all subsequently selected support features are not correlated to any of the previously selected support feature. Last but not least, the top scoring features that satisfy the constraint in (4.5) then form the support features, hence $\max_{\boldsymbol{\delta},\mathbf{f}_j \notin \mathbf{X}_{\mathcal{AF}.\mathcal{G}_z}} \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta})\|^2$ holds. Inductively, it becomes possible to conclude that the proposed CRM algorithm can solve (4.11). This completes the proof.

The next theorem indicates that the proposed GDM can globally converge and exhibits the non-monotonic property for feature selection.

Theorem 3. *For each iteration of Algorithm 2, suppose that the reduced minimax subproblem (4.8) can be globally solved and the most violated constraint selection (4.11) can be solved with the most informative feature selected, Algorithm 2 terminates after a finite number of iterations.*

Proof: The convergence of the proposed algorithm can be measured by the gap difference between series $\{\beta^k\}$ and $\{\varphi^k\}$ (please refer to Theorem 2). Further, after a finite number of iterations, the difference of objective values from adjacent iterations is close to 0. Assuming that at the k^{th} iteration, there is no fresh updates on $\mathcal{C}^k$ (*i.e.*, $\boldsymbol{\delta}^{k+1} = \arg\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}^k)$ and $\mathcal{C}^{k+1} = \mathcal{C}^k$). Therefore, one can prove that, in this case, ($\boldsymbol{\alpha}^k$,β^k) is the optimal solution pair of (4.7). Further, since $\mathcal{C}^{k+1} = \mathcal{C}^k$, in Algorithm 2, there will not be any change to $\boldsymbol{\alpha}$ as well (*i.e.*, $\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k$). Then, we have $g_{\boldsymbol{\delta}^{k+1}}(\boldsymbol{\alpha}^k) = \max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}^k) = \max_{\boldsymbol{\delta}\in\mathcal{C}^k} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}^k) = \max_{1\le i\le k} g_{\boldsymbol{\delta}^i}(\boldsymbol{\alpha}^k) = \beta^k$, and $\varphi^k \le \min_{1\le j\le k} g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j) \le \beta^k$. In addition, from Theorem 2, $\beta^k \le \theta \le \varphi^k$ holds only when $\beta^k = \theta = \varphi^k$, providing the solver ($\boldsymbol{\alpha}^k$,$\beta^k$) as solution to Algorithm 2. This completes the proof.

In general, the *Cutting Plane Algorithm* typically converges to robust optimal solution within tens of iterations under the worst case analysis (i.e., finding the most violated constraint $\boldsymbol{\delta}^t$), where notable performances on many real applications have been reported [155].

Table 4.1: Complexity analysis for each iteration in GDM.

Sub-procedure		**Complexity**
CRM	Computing Feature Score $\mathbf{s}$	$O(mS)$
	Sorting $\lvert s_j \rvert$'s	$O(m \log m)$
	Grouping Features	$O(\mathcal{K}_b mn)$
MKL		$O(T\mathcal{K}_b n)$

4.3.10 Complexity Analysis

In GDM, the most violated $\boldsymbol{\delta}$ is obtained via the CRM algorithm, where the m number of features are firstly sorted based on the feature score metric $|s_j|$ in GDM. Consequently, $\mathcal{K}_b$ number of most informative features with correlation consideration (i.e., w.r.t. the other predictive features) are identified. For T iterations, there will be $T \times \mathcal{K}_b$ SFs at most, which is the worst case to consider in MKL. Nevertheless, as previously discussed, the cutting plane strategy requires a small T for convergence to happen – a cap of ten iterations is used in the experimental study on binary classification, thus T is not a crucial term in the complexity. For the sake of brevity, the detailed complexity analysis on the two iterative steps of the proposed method is given in Table 4.1, where S indicates the number of support vectors in the SVM. Note that for the Feature Grouping stage of CRM, the complexity $O(\mathcal{K}_b mn)$ is for the worst case, however, due to the phenomena of "sparse correlation" highlighted in the Chapter 2, the true complexity is much less than $O(\mathcal{K}_b mn)$. To conclude, GDM is very efficient for the real-world data with "sparse correlation".

4.4 Experimental Study

In this section, the experimental study on the proposed GDM-PCC and GDM-SU are presented together with several state-of-the-art feature selection methods, including: 1) mRMR[6] [161], 2) FCBF[7] [162], 3) RCFS [7], 4) SVM-RFE [74] and 5) ℓ_1-SVM[8] [136]. In addition, some state-of-the-art feature grouping methods are also considered here to

[6]http://penglab.janelia.org/proj/mRMR.
[7]http://www.cs.man.ac.uk/~gbrown/fstoolbox.
[8]http://www.csie.ntu.edu.tw/~cjlin/liblinear.

assess the performance efficacies of GDM. These include: 6) OSCAR [86], 7) ncFGS & ncTFGS [127] and 8) GFlasso [76]. To provide insights to the contributions of incorporating correlation constraints, $\tau = 0$ is configured to arrive at GDM$_{\tau=0}$. This forms the baseline where no differentiations between support and affiliated features are made (i.e., features with highest $|s_j|$'s are preferred).

4.4.1 Experimental Setup

Among the feature selection methods considered here, mRMR, FCBF and RCFS represent filter methods, SVM-RFE is a representative of the wrapper method, while ℓ_1-SVM belongs to the family of embedded method. For the feature grouping methods, OSCAR, ncFGS & ncTFGS and GFlasso all operate based on the strategy of pruning the covariance/correlation matrix. The configurations of all the methods considered are set to be consistent to those used in the respective articles reported, and implemented in C++. Moreover, to facilitate fair comparisons, the standard SVM classifier is used as the underlying classifier. In GDM, the parameter C of the standard SVM is set to '1', while in ℓ_1-SVM, C is unique and vary with the number of features selected. In the experimental study, the correlation parameter is set as $\tau = 0.3$ for GDM-PCC while $\tau = 0.4$ for GDM-SU, and for the mutual information calculation in GDM-SU, we use 10% of the feature value range as the estimator. Further, to show the results of different numbers of selected features, $\mathcal{K}_b$ is set as natural number (e.g., 1, 2, ..., 10). All experiments are conducted on the PC with Intel® Xeon® CPU E5-2695 v2 (2.4 GHz, 2 processors) and 128GB memory under Windows Server® 2012 R2 Standard.

4.4.2 Results on Synthetic Dataset

To illustrate the mechanisms of the proposed GDM, the study begins with a synthetic dataset, where the ground truth support-affiliated relationships are known in advance. Correspondingly, the aim is to verify if our proposed method is able to discover the feature groups and how it fares against the existing state-of-the-art feature grouping methods. The training set comprises 2,048 observations, each having 10,000 features. There are 12 predefined informative features which have been further expanded as 12 feature groups with variant size, as depicted in Figure 4.4.a. Each support feature has 0 to 5 affiliated

Table 4.2: Results on synthetic dataset of various methods. **Success Hits** stands for the completeness in identifying the features. It measures the matching degree for feature grouping methods (i.e., $\frac{\#\text{CORRECT}}{\#\text{ALL}} \times 100\%$) whilst taking the form of "# correct SF/# correct AF/# incorrect feature" for feature selection methods. The **Training Time** is in reported seconds, wherein the deviation below 1 second is reported.

	Feature Grouping Methods				
	GDM	OSCAR	ncFGS	ncTFGS	GFlasso
Success Hits	**86.8%**	50.0%	60.5%	52.6%	71.1%
Training Time	0.85±0.13	130.68	455.34	456.62	132.49

(a) Performances of Feature Grouping Methods

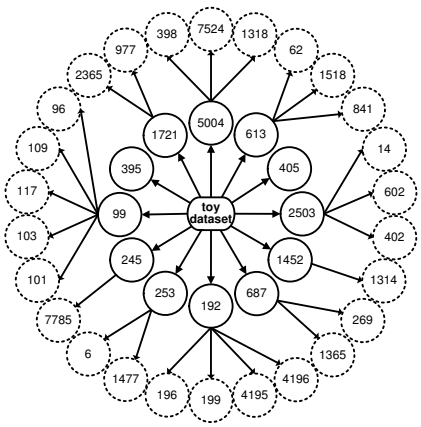
	Feature Selection Methods					
	$\text{GDM}_{\tau=0}$	mRMR	FCBF	RCFS	SVM-RFE	ℓ_1-SVM
Success Hits	**9/3/0**	2/3/7	6/1/5	7/2/3	8/3/1	**10/2/0**
Training Time	**0.31±0.16**	2.28±0.25	4.56±0.52	101.95	33.73±0.98	**0.33±0.07**

(b) Performances of Feature Selection Methods

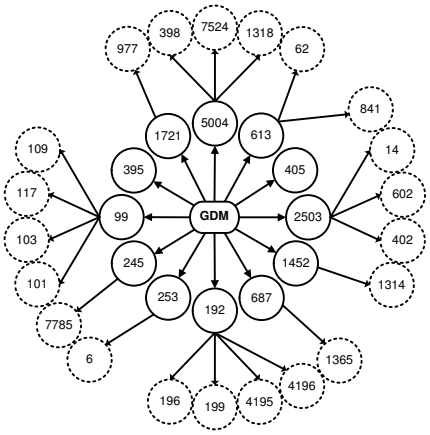
features as children, while the others then form the noisy features. The predictive ability of each group is configured to follow a normal distribution $\mathcal{N}(0, 1)$. The pseudo algorithm of generating this dataset is provided in the Section ii of the Appendix, and since the linear correlation is used to generate the feature group, the GDM-PCC is adopted in the experiment. Thus, in what follows, GDM indicates GDM-PCC in this section.

The experimental results obtained by the various feature grouping methods and feature selection methods on the synthetic dataset are tabulated in Table 4.2. In particular, success hits and training time are the performance metrics used to assess the algorithms under consideration as summarized in the table. *Success Hits* provides a measure on the completeness of a feature grouping method or feature selection method in correctly identifying all the core features. *Training Time*, on the other hand, gives the wall-clock time incurred to train a learning model. From the results in Table 4.2, both the wrapper (i.e., SVM-RFE) and embedded methods (i.e., GDM, ℓ_1-SVM) have been observed to attain competitive performances on both metrics for feature selection methods.

With the correlation constraint in (4.5) disabled by setting $\tau = 0$, $\text{GDM}_{\tau=0}$ is observed to achieve performances that are close to the ℓ_1-SVM. This is unsurprising due to



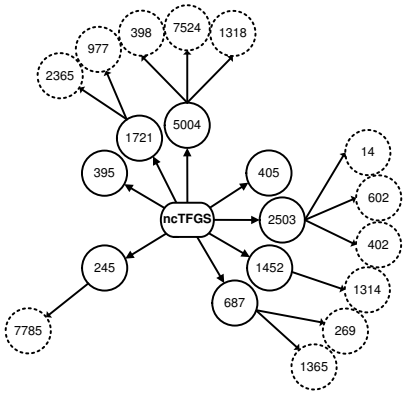
4.4.a: Ground truth



4.4.b: GDM

4.4.c: OSCAR

4.4.d: ncFGS



4.4.e: ncTFGS

4.4.f: GFlasso

Figure 4.4: Feature Group Structures generated by the various feature grouping methods.

the similar sparse SVM strategy used in both algorithms. Filter methods such as mRMR and FCBF suffered the worst performances. Nevertheless, RCFS, though achieved competitive result, incurred a learning time of 102 seconds, which is 300+ times more than embedded feature selection method, i.e., $\text{GDM}_{\tau=0}$ and ℓ_1-SVM.

The feature group structures generated by the various feature grouping methods are then depicted in Figure 4.4.b – Figure 4.4.f. Visually, Figure 4.4.b which is the group structure produced by the proposed GDM, is noted to match the ground-truth feature groups of Figure 4.4.a most closely than all the others, i.e., (Figure 4.4.c – Figure 4.4.f). To assert this quantitatively, *Success Hits* is used to measure the obtained feature group structures of the various feature grouping methods (i.e., matching degree for feature grouping methods, $\frac{\#\text{CORRECT}}{\#\text{ALL}} \times 100\%$)), among which GDM ranked at the top, with a value of 86.84% (i.e., 33/38 of ground truth features). GFlasso managed to uncover 27/38 of the ground truth features, while all the other feature grouping methods only identified less than 23 of the ground truth features. For the sake of illustration, the important features that have been identified by the feature grouping methods are depicted in Figure 4.4.

4.4.3 Results on Real-world Datasets

To assess the performance of GDM on real world settings, here a study is presented on a range of big dimensional datasets from diverse domains. The first is the 20 newsgroup dataset, which has been size-balanced for binary text classification with each class comprising 10 groups and labelled here as `news20.binary`[9]. The second is the `kdd2010`[9] challenge dataset used in the educational data mining competition. The aim of the competition is to provide predictions on the "correct first attempt" for a subset of "steps". The third is the spam web page data `webspam`[10], which is collected by "Webb Spam Corpus 2006" with adequate number of spam pages for the purpose of spam detection. One biology data considered here is the `psoriasis` dataset comprising Single-Nucleotide Polymorphisms (SNPs) as the features. This data is collected from a collaborative association study of psoriasis (CASP)[11] to identify susceptibility pathways and important genes [178]. As SNPs data is very dense, this makes feature grouping and selection a challenging and non-trivial task.

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets
[10] http://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html
[11] http://www.sph.umich.edu/csg/abecasis/casp

Table 4.3: Characteristics of the real-world datasets considered.

Dataset	# Features	# Training	# Positive	# Negative	# Testing
news20.binary	1,355,191	9,996	6,000	3,996	10,000
kdd2010	**29,890,095**	**19,264,097**	16,579,660	2,684,437	748,401
webspam	16,609,143	280,000	169,786	110,214	70,000
psoriasis	529,651	2,176	1,131	1,045	545

Dataset	# Nonzeros	Density	Size on disk
news20.binary	3,584,383	2.646e-04	133.52MB
kdd2010	585,609,985	1.017e-06	4.96GB
webspam	1,044,482,369	2.245e-04	**23.31**GB
psoriasis	**1,424,895,775**	**0.989**	11.67GB

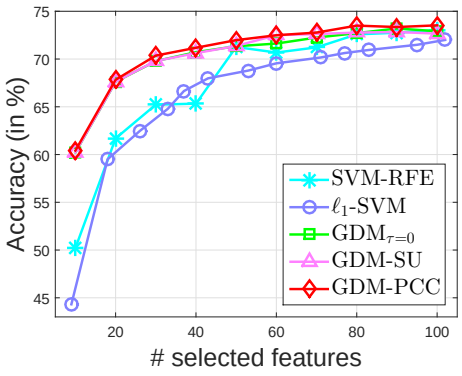
To proceed with our study on the webspam and psoriasis datasets, 80% of the entire observations are randomly selected to form the training set, and the rest 20% are kept for testing. For news20.binary, the training and testing set are set to be equal in size due to the sparseness of the dataset, so as to better preserve the original feature space. Further, detailed information on the datasets is listed in Table 4.3, wherein the bold font indicates the core challenge of each dataset considered, e.g., the challenge of big dimensionality in kdd2010 with nearly 30 million features, high density characteristic of psoriasis and enormous storage requirement of webspam, etc.

To adapt with the feature selection task, SFs are used to represent GDM series methods, *e.g.,* GDM-PCC or GDM-SU and GDM$_{\tau=0}$. Further, to evaluate the feature selection performances, *classification accuracy*, *training time complexity* and *redundancy rate*[12] [81] are considered.

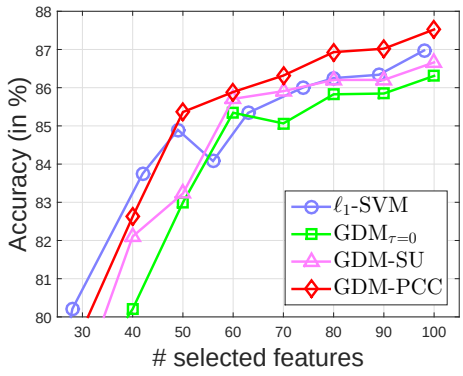
4.4.3.1 Accuracy Results

Figure 4.5 summarizes the accuracy performance attained by the methods considered including SVM-RFE, ℓ_1-SVM, GDM$_{\tau=0}$, GDMs (GDMs refers to both GDM-SU and GDM-PCC in this subsection). Note that, all the other feature grouping methods as well as the filter feature selection methods have been observed to be inadequate for handling such high dimensionality considered, hence only results of the wrapper and embedded methods are reported in the figure. Furthermore, SVM-RFE consumed a large number

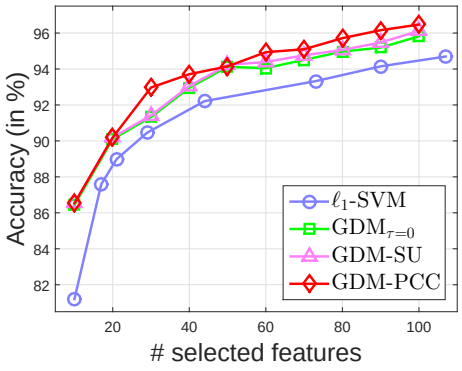
[12]Redundancy rate assesses the averaged correlation among all the selected feature pairs.



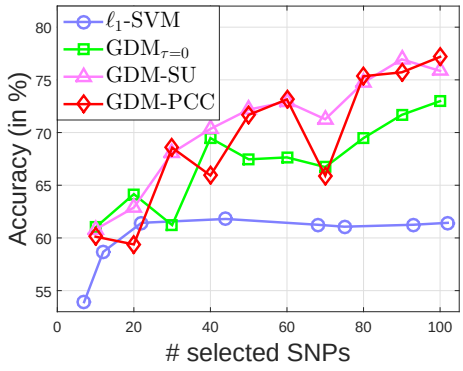
4.5.a: `news20.binary`



4.5.b: `kdd2010`



4.5.c: `webspam`

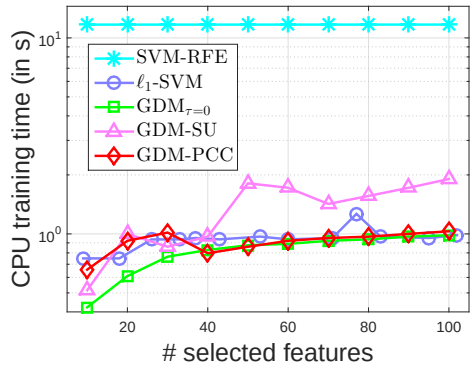


4.5.d: `psoriasis`

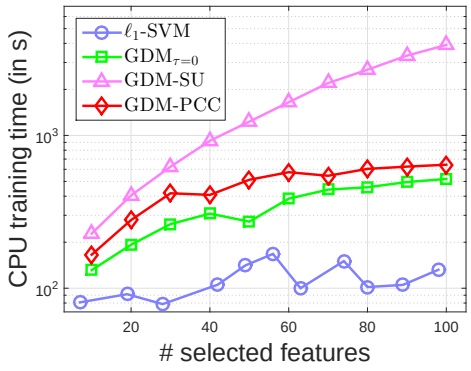
Figure 4.5: Testing accuracy (in %) on real-world datasets.

of inner SVM evaluations, thus it fails to converge well on the three larger datasets under the limited training budget available.

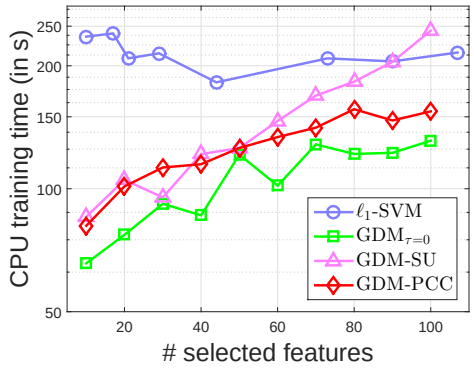
Overall, GDM attains very high accuracy improvements over the other methods on the real world datasets. From Figure 4.5, one can also observe that both GDM-PCC and GDM-SU fare significantly better than GDM$_{\tau=0}$ on the larger datasets. This improvement can be attributed to the benefits brought about by the feature correlation constraints considered in GDM, i.e., identifying the SFs and AFs, since the only disparity between GDMs and GDM$_{\tau=0}$ lies in the lack of differentiations between SFs and AFs in the latter. While for `news20.binary`, since the informative features are mostly uncorre-



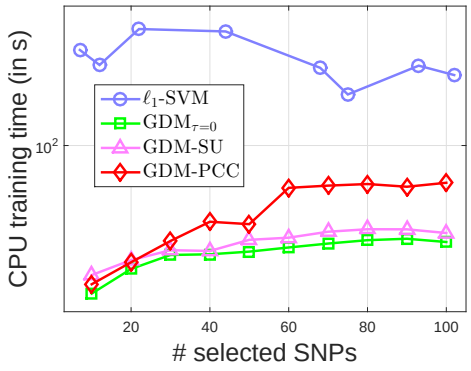
4.6.a: `news20.binary`



4.6.b: `kdd2010`



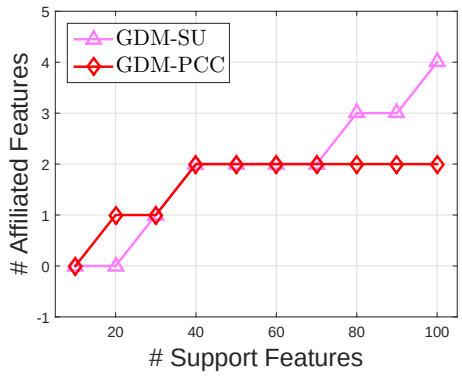
4.6.c: `webspam`



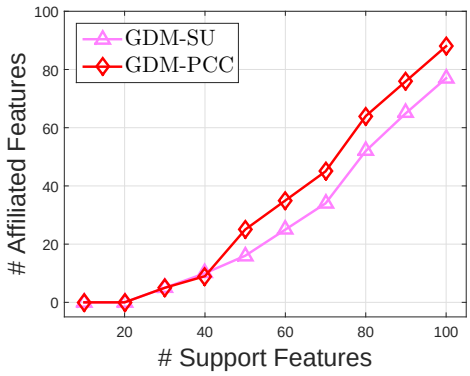
4.6.d: `psoriasis`

Figure 4.6: Training time (in seconds) on real-world datasets (in logarithmic scale, averaged from 5 runs).

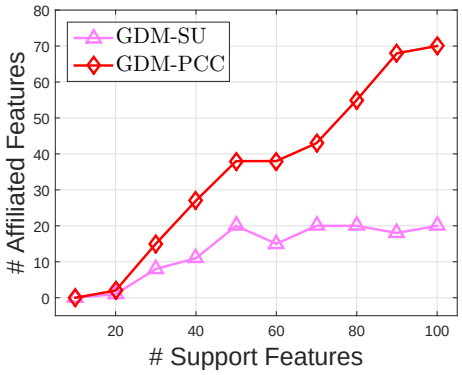
lated, only a small number of feature groups are identified by both GDM methods, hence similar accuracies are reported by GDMs and GDM$_{\tau=0}$. Besides, GDM-PCC performs slightly better than GDM-SU on three of the datasets, while on the `psoriasis` biological data, the nonlinear correlation measure of GDM-SU exhibited improved and more stable accuracy, especially when the number of selected SFs is small. Moreover, both GDMs and GDM$_{\tau=0}$ showcase statistically significant improvements in accuracy over the ℓ_1-SVM, however, relatively high variance is displayed in the prediction accuracies reported w.r.t. the increasing selected SNPs, which is mainly due to the *high dimension*



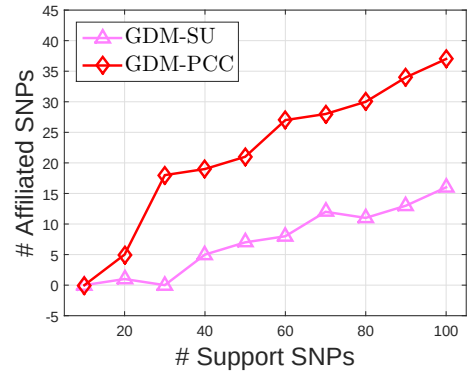
4.7.a: `news20.binary`



4.7.b: `kdd2010`



4.7.c: `webspam`



4.7.d: `psoriasis`

Figure 4.7: Number of AFs selected regarding to number of SFs by GDM-PCC and GDM-SU.

small sample size characteristic of the `psoriasis` dataset that will be illustrated further in a subsection of further study (see Section 5.2).

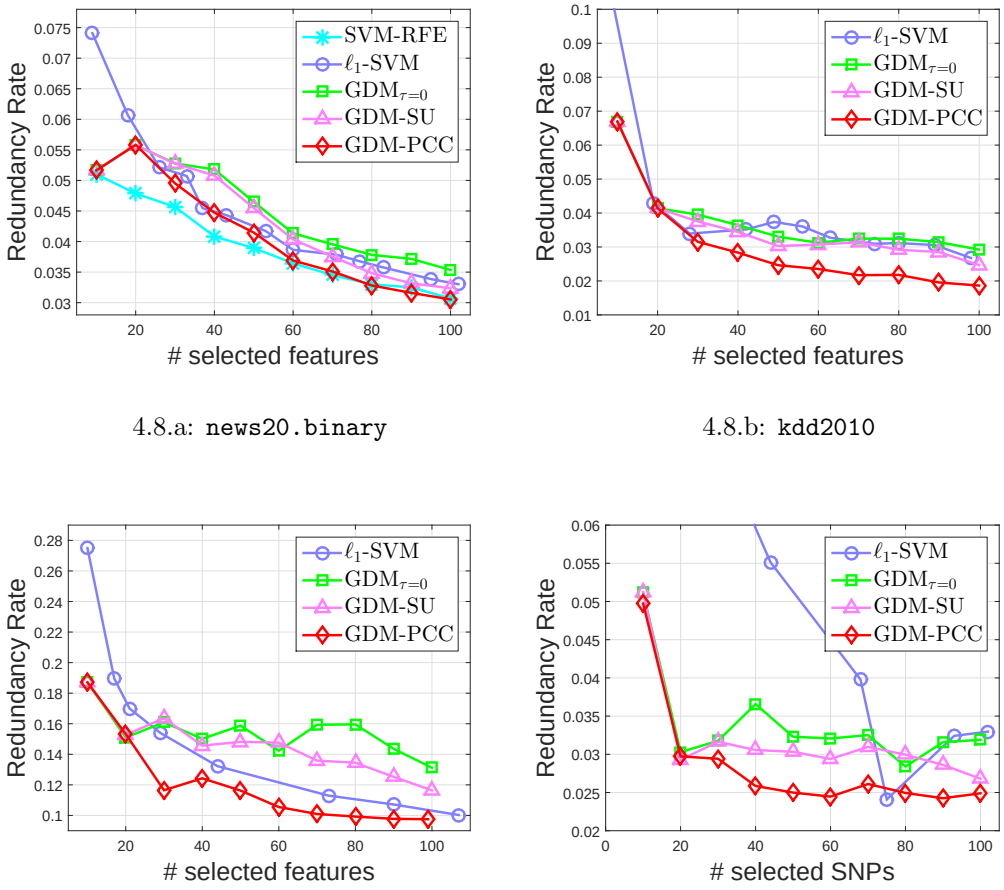
4.4.3.2 Training Time Results

Figure 4.6 further summarizes the training cost incurred, wherein embedded methods are noted to be more efficient on the `news20.binary` dataset when compared to the wrapper method, SVM-RFE. Moreover, Figure 4.6.b shows that, on the highly sparse `kdd2010` dataset (see Table 4.3), ℓ_1-SVM was able to take advantage on the sparseness character-

istics of the kdd2010 dataset to achieve the shortest training time observed. However, on denser datasets, i.e., psoriasis, ℓ_1-SVM did not fare well and in fact incurred large training costs due to the inadequate management of memory resources. When no differentiations between SF and AF are considered, $\text{GDM}_{\tau=0}$ displays remarkable training efficiency compared to GDM-PCC and GDM-SU, as shown in Figure 4.6.b and Figure 4.6.c, with some tradeoff in prediction accuracies. Most importantly, $\text{GDM}_{\tau=0}$ could possibly satisfy the real-time requirement of some applications if some form of parallel computing is used. Despite having to handle the massive number of feature correlation computations involving big dimensional dataset, GDM-PCC reported a comparable or smaller training time costs than ℓ_1-SVM. Further, referring to the number of selected SFs and AFs reported in Figure 4.7, it can be observed that when the quantity of AFs in GDM-PCC and GDM-SU are similar, and GDM-SU incurs a higher computational time than GDM-PCC (e.g., as observed on the news20.binary and kdd2010 dasasets). However, on the dense dataset psoriasis, GDM-SU exhibits higher efficiency with a smaller identified feature groups (note that for webspam, GDM-PCC may take advantage of the sparseness in the dataset to reduce the cost of PCC calculation).

4.4.3.3 Redundant Rate Results

Last but not least, in Figure 4.8, the redundant rate attained by various methods are depicted, wherein GDM-PCC achieves a low rate in most cases. Compared to embedded methods, both GDM-SU and GDM-PCC outperform the ℓ_1-SVM in terms of redundancy reduction, while exhibiting improved accuracy performance at the same time. Moreover, this implies that the SFs selected by GDM-SU or GDM-PCC approximately form a *good feature set* [78]. Specifically, from Figure 4.8.a, SVM-RFE also reported low redundancy rate. However, tracing back to Figure 4.5.a, SVM-RFE did not fare so well on the accuracy performance metric. It appears that the redundancy of SVM-RFE suffers from the presence of irrelevant features. Overall, considering both the results in Figure 4.5 and Figure 4.8, both GDM-PCC and GDM-SU are noted to attain low redundancy rate and superior accuracy performance simultaneously on the real-world datasets and relative to all the feature selection methods considered.



4.8.a: `news20.binary`

4.8.b: `kdd2010`

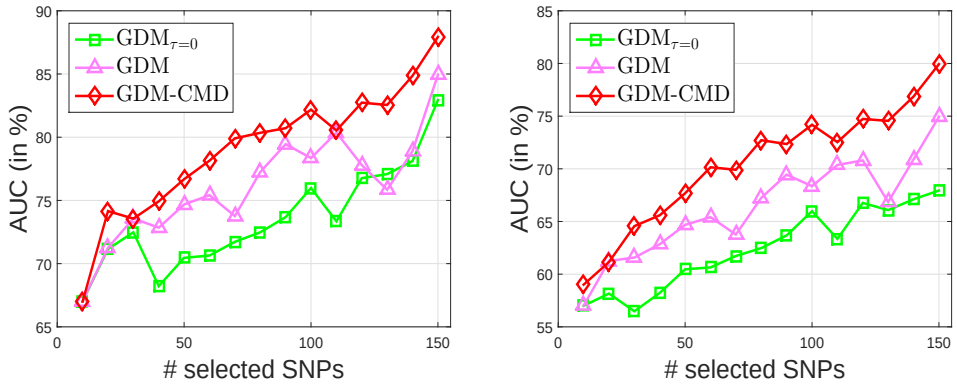
4.8.c: `webspam`

4.8.d: `psoriasis`

Figure 4.8: Redundant rate on real-world datasets.

4.4.4 Results on Psoriasis Dataset Using Specific Correlation Constraint

For the purpose of selecting as fewer features as possible, hereby only SFs are sought to identify the major information in the data. Moreover, AUC, as a notable accuracy measure, denotes the Area Under the ROC (Receiver Operating Characteristics) Curve in statistics. This metric has been commonly used in feature selection methods when dealing with bioinformatics problems [179]. Notably, Ling *et al.* showed the statistical consistency of AUC both mathematically and empirically, and concluded that it serves as a more discriminating measure than traditional classification accuracy [180], especially



4.9.a: `psoriasis` 4.9.b: `psoriasis-larger`

Figure 4.9: AUC results of various methods.

on some imbalance class distribution tasks. Thus, in this subsection, AUC is adopted as the performance metric for the SNPs datasets to test the specifically proposed correlation constraint. The GDM that adopts the constraint of centimorgan distance is termed as GDM-CMD in this experiment. Further, the experimental setup is configured as follows: $\tau = 0.4$ for both GDM and GDM-CMD, while β=$2m$ for GDM-CMD. Besides, C is configured to 1 for all methods.

The SNPs of interest in this work are the Psoriasis datasets. Psoriasis is a chronic inflammatory disease that affects the skin, nails and joints. It occurs when the immune system mistakes the skin cells as a pathogen, and sends out faulty signals that speed up the growth cycle of skin cells. The psoriasis dataset that used in the previous subsection is considered as a core subset. In this subsection, besides the subset, the complete dataset which comprises 1,530,904 SNPs is also included. Note that this dataset shares the identical general information as described in Table 4.3 with its subset, hence the two datasets can be regarded as unique expressions under similar conditions. However, this dataset is extremely large, which occupies 35.2 GB on hard disk. For simplicity, hereinafter, the two datasets will be denoted as `psoriasis` and `psoriasis-larger`, respectively.

The AUC performance results on the Psoriasis SNPs datasets are then summarized in Figure 4.9. Relative to GDM$_{\tau=0}$ and GDM, GDM-CMD exhibits a significant im-

provement under the measurement of AUC. However, $\text{GDM}_{\tau=0}$ did not perform as well as the other two methods, since it does not take feature correlation into consideration during the search. Compared with GDM, GDM-CMD outperforms GDM in terms of the new constraint with prior knowledge embedded towards advanced redundancy reduction. This also supports the reality, that is, if the two correlated SNPs are distant, there may be high probability that they belong to different gene, consequently it is safe to avoid the affiliated relationship for them. Note that, GDM-CMD achieves a 88% AUC on `psoriasis`, when the number of selected features is only 150, while 80% on `psoriasis-larger` out of the original feature set of 1.5 million SNPs.

4.5 Summary

Today, modern databases with Big Dimensionality are experiencing a growing trend. State-of-the-art approaches that require the calculations of pairwise feature correlations in their algorithmic designs have not coped well on such database, since the computation burden of m^2 is often impractical. In this chapter, the observations from several real-world databases have established that an extremely small portion of the feature pairs contribute significantly to the underlying feature interactions, i.e., there is a presence of sparse correlations, and there exists feature groups that are highly correlated. Taking the cue, this research thus embarked on a comprehensive study on potential correlated informative features or feature groups using the concepts of support feature and affiliated feature, to fill in the research gap that has been identified. In particular, the proposed GDM embeds an explicit incorporation of both linear and nonlinear correlation measures as constraints in the learning model to filter out large number of non-contributing correlations that could otherwise confuse a classifier, while identifying the correlated and informative feature groups. Notably, the affiliated features are constructed in the proposed method without any additional cost, since they are generated along with the support features.

Moreover, extensive empirical studies have been reported on both synthetic and several real-world datasets to reveal the superior prediction accuracy of GDM through the identified SFs. However, for a fair comparison, the performance of the AFs cannot be demonstrate through same measurements, hence some feature redundancies via the identification of AFs has been studied in next chapter, where one can observe that the AFs

or feature groups are useful for enhanced interpretation of the learning tasks. Note that, as the method is specially designed for big dimensional data, one drawback observed from the experimental results shows that the method may not be very contributing when the dimension of data is not enough. Besides, the proposed GDM is powerless with the unbalanced data in terms of correlation, e.g., the extreme case where there is only one real feature, while other features are the duplicates of this feature. Furthermore, with the centimorgan distance improving the performance of the GDM, the proposed framework is demonstrated to be flexible in adapting specific correlation constraint generated. Last but not least, both the theoretical analysis and empirical studies verified the high efficacies of the proposed methodology, which makes trillion correlations feasible when dealing with big dimensional data.

Chapter 5

Robustness and Further Discussions on the GDM Framework

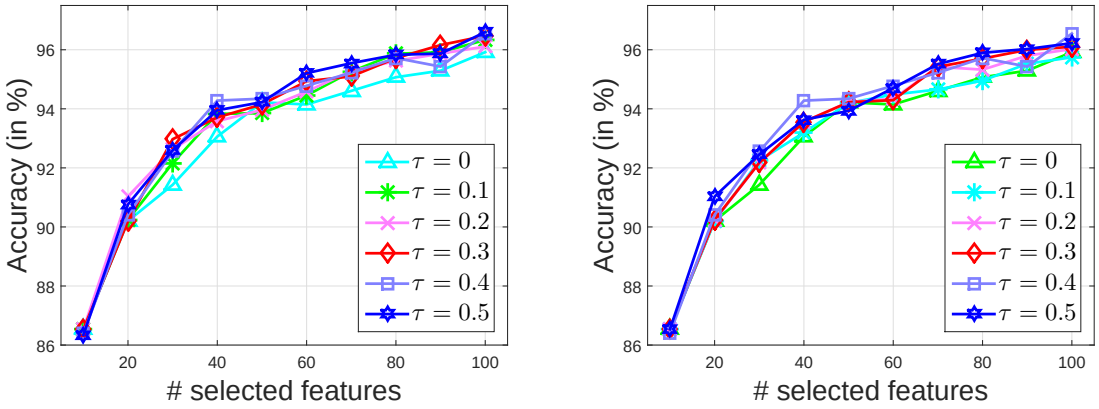
In this chapter, the sensitivity analysis of the parameters in GDM is firstly discussed from the perspective of feature correlation to show the stability and robustness of the proposed GDM framework. Further, to identify robust informative features with minimal sampling bias, the embedding of the V-fold cross validation in the GDM is considered to seek for features that exhibit stable or consistent performance on multiple data folds. Consequently, the proposed GDM framework on one class learning problem is discussed in terms of reducing computational cost on big dimensional data. Last but not least, the benefits of affiliated features and selected feature groups are presented from different tasks with different techniques and various real-world datasets.

5.1 Sensitivity Analysis

Since the GDM framework has shown promising results on the real-world datasets of interest, there emerges a natural question regarding the sensitivity of GDM with respect to the parameters. This section thus serves the purpose of answering some of these question. Specifically, the pairwise correlation threshold τ and the estimated interval for mutual information (i.e., used in the non-linear correlation constraint) are analysed.

5.1.1 Sensitivity Analysis of Parameter τ

In this subsection, we experimentally study the effects of τ for GDM-PCC and GDM-SU on `webspam` dataset. In particular, the results of GDM-PCC for different values of τ at {0,



5.1.a: GDM-PCC

5.1.b: GDM-SU

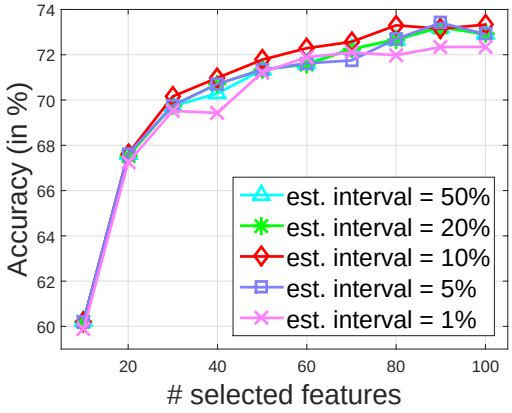
Figure 5.1: Testing accuracy of GDM methods with different value of τ on the `webspam` dataset.

0.1, 0.2, 0.3, 0.4, 0.5} are summarized in Figure 5.1, where we observe that the curves are close to one another in both figures. Thus it can be concluded that they exhibit similar trends across different value of τ. Particularly, GDM-PCC with $\tau = \{0.2, 0.3, 0.4, 0.5\}$ and GDM-SU with $\tau = \{0.2, 0.3, 0.4, 0.5\}$ share very similar performance. From Figure 5.1.a, it can be observed that the best result of 96.16% in accuracy is reported for GDM-PCC with $\tau = 0.3$, where 90 support features have been selected. On the other hand, the best accuracy of 94.28% is found for GDM-SU with $\tau = 0.4$ with 40 selected support features, as can also be observed in Figure 5.1.b.

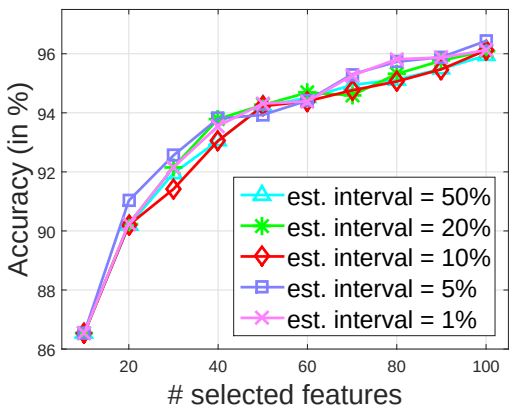
Thus, from the results obtained, GDM methods are found to be not strongly sensitive to the realistic range of parameter τ. However, in real applications, where prior knowledge is available, suitable value of τ can be used to improve the testing result to define accurate feature groups.

5.1.2 Sensitivity Analysis of Estimation Interval Selection in Mutual Information

In this subsection, we study the effects of the estimation intervals in the mutual information criterion on GDM-SU performance accuracy. Here, we consider the feature



5.2.a: `news20.binary`



5.2.b: `webspam`

Figure 5.2: Testing accuracy of GDM-SU on different estimated interval in mutual information.

distributions and arrive at suitable feature range. Thus, we take 50% (2 categories), 20%, 10%, 5%, 1% of the feature range as the estimation interval used in the experimental study on the real-world datasets, particularly, `news20.binary` and `webspam`. The corresponding test results are summarized in Figure 5.2, where it can be observed that the impact of the estimation interval on GDM-SU is insignificant (the chosen interval in the main manuscript, i.e., 10%, is not dominating as observed from the curves). Even on the extreme case of 100 intervals (i.e., 1%), the accuracy is noted to remain resilient.

To summarize, GDM is found to be insensitive to the estimator interval used, since the accuracy result mainly depends on the feature score, which is determined in the feature selection part of GDM.

5.2 Robust Feature Selection – GDM with Embedded Cross Validation

As a result observed from Figure 4.5, the accuracies of GDM-PCC grow smoothly on three of the real-world datasets, however, shake sharply on the `psoriasis` dataset, which is possibly due to the *high dimension small sample size* characteristic of the data. In practice, when dealing with big dimensional data, the feature size sometimes far exceeds

Input: Dataset $\mathcal{D}(\mathbf{X}, \mathbf{y})$, $\mathcal{K}_b$, τ and number of fold V.
Output: Index sets of $\mathcal{SF}$ and $\mathcal{AF}$.
Initialize $\boldsymbol{\alpha} = \mathbf{1}/n, \mathcal{SF} = \varnothing$ and $\mathcal{AF} = \varnothing$, and randomly split the training set into V equal partitions $\mathcal{D}_v$.
for $t = 1$ **to** T **do**
  Compute feature score vector $\mathbf{s}$ and sort $|s_j|$ in descending order, record the feature ranking list as $\mathcal{E}$
  for $v = 1$ **to** V **do**
    Call $\boldsymbol{\delta}_v^t = \text{CRM}(\mathcal{D}_v, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}_v, \mathcal{Q}_v)$
  end for
  1: Set '1' for the indices of top $\mathcal{K}_b$ features from $\sum_v \boldsymbol{\delta}_v^t$ in $\boldsymbol{\delta}^t$
  2: Archive $\mathcal{AF}$ from each $\mathcal{AF}_v$ that corresponding to $\mathcal{SF}$
  3: Set $\Lambda = \Lambda \cup \{\boldsymbol{\delta}^t\}$ while updating $\boldsymbol{\alpha}^{t+1}$
end for

Algorithm 5: GDM with Embedded Cross-Validation

the number of data samples by many orders of magnitude [181, 182]. In such cases, it is typical that any slight variations in the training data often produces radical changes in the selected feature models. In order to identify an intrinsic set of features that represent the entire data or the original feature space, here the feature selection results of multiple feature selectors are aggregated by means of voting or averaging, in the spirit of the ensemble feature selection strategy [181]. With a consensus made using diverse feature selectors, possible biases caused by the inconsistent distributions of the data can be reduced. For example, Figure 5.3 showcases the process of making the consensus, where features #2, #5, #11, #13 and #25 have been voted to the "robust feature subset".

In achieving such a goal, the universal support feature set $\boldsymbol{\delta}^{uni}$ is elected, which is formulated as

$$\min_{\mathbf{w}^v, \boldsymbol{\delta}^{uni}} \sum_v \text{GDM}(\mathbf{w}^v, \boldsymbol{\delta}^{uni}, \mathcal{D}^v), \tag{5.1}$$

wherein the superscript $\cdot^v$ in the function denotes the v^{th} feature selector, hence $\mathbf{w}^v$ represents the corresponding weight vector and $\mathcal{D}^v$ the data subset. Further, a universal feature set $\boldsymbol{\delta}^{uni}$ is enforced to identify only single set of robust SFs kept in testing (*i.e.*, rather than one set per cross validation run). To solve problem (5.1), the traditional V-fold cross-validation is embedded within the GDM. Consequently, this framework is labelled here as Embedded Cross Validation (ECV), whose algorithm is summarized in

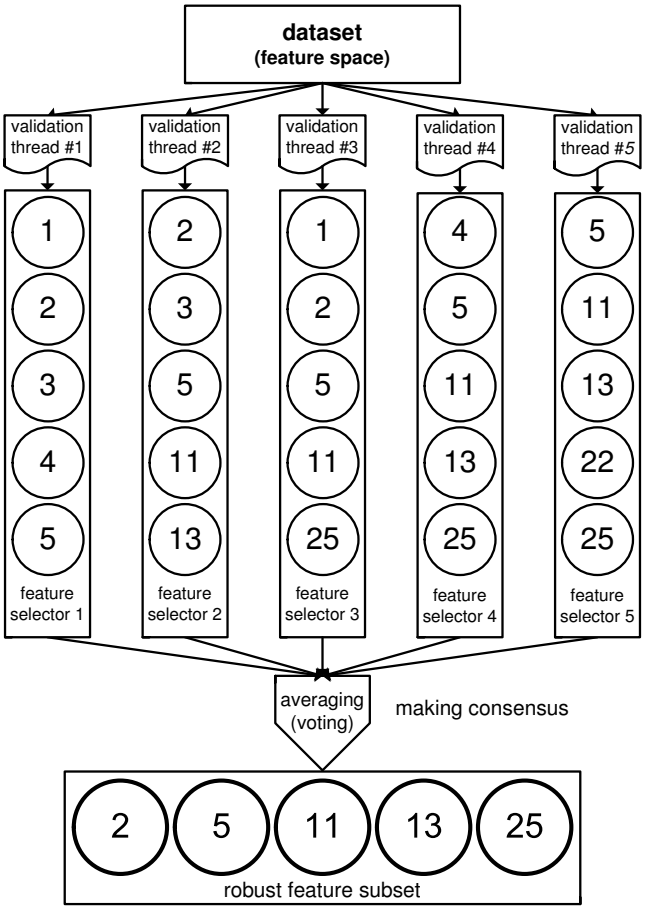


Figure 5.3: Concept and principle of robust feature selection.

Algorithm 5. Further, with such adoption, the complexity only involves the CRM V times, while exhibiting only one robust feature set.

Thus this data is trained using ECV, where $V = 5$ folds are adopt in the procedure. One can observe from Figure 5.4 that ECV not only outperforms GDM in terms of prediction accuracy, it also reports a significantly more stable results. This underlines the benefits of embedding the cross validation with GDM to arrive at ECV that helps alleviate the bias of such data distribution, so as to converge to robust features. However, ECV exhibits increasing training time costs w.r.t. the selected features due to the embedded cross-validation scheme of GDM but less than V-fold of the time as discussed. Thus, it is worth emphasizing that the aggregation of performances by diverse feature selector is helpful for improving the robustness of the selected features, especially in *high dimension*

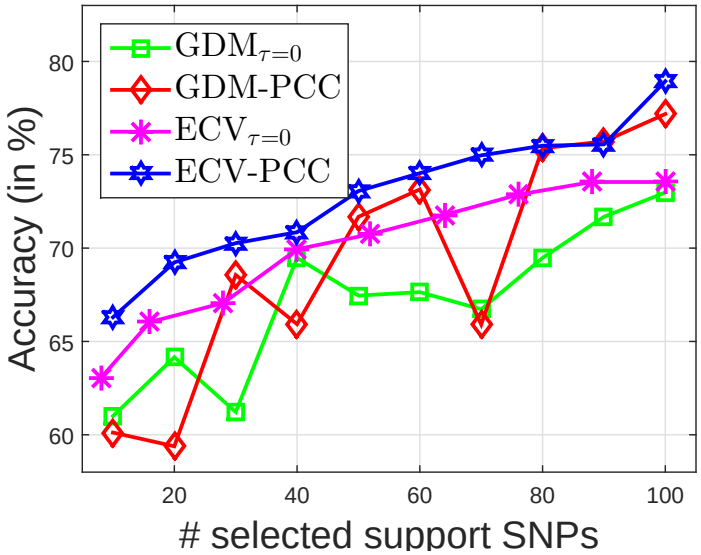


Figure 5.4: Experimental results on `psoriasis` for ECV.

small sample size problem settings [59, 183], wherein different feature subsets are often reported to yield similar performances.

5.3 Further study on One-Class Learning

Beside standard binary classification task, one-class machine learning (i.e., novelty detection, oppositely) has also attracted numerous interests in the literature, whose major task is similar to a binary classification problem with the exception that it differentiates known objects (i.e., normal patterns) from abnormal objects with different labels or even without labels, while preserving the distribution of the known objects as one-class. Obviously, locating such robust decision boundary which covers the normal patterns only is the primary challenge of one-class problem. Moreover, embedded feature selection strongly improves one-class problem in efficiency since in SVM training, subset of data instead of entire training set is used. However, such framework cannot be simply adapted by a clustering-based method, wherein a small perturbation can lead to very different performance.

In this section, the generality of the proposed GDM framework for solving *One-Class Learning* problems is further illustrated. Particularly, many applications, including fraud

detection and novelty detection, are commonly seen as one-class learning problems. Under such a problem setting, an expected classification is taken to differentiate between known objects (i.e., the target class) from unknown objects (i.e., outlier, abnormal observation), while preserving the corresponding distribution of the known. Therefore, data with single class label can be used to assess whether a learning machine is able to properly preserve the boundary of the learned class. In what follows, the proposed GDM framework is shown to be capable of readily accommodating one-class problem with ease by holding the label information of each observation (i.e., $y_i = 1$ for all known objects) in the proposed sparse SVM formulation,

$$\begin{aligned} \min_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \min_{\mathbf{w},\gamma,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 - \gamma + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \\ s.t. \quad & \mathbf{w}'(\mathbf{x}_i \odot \boldsymbol{\delta}) \geq \gamma - \xi_i \quad i = 1, \cdots, n. \end{aligned} \tag{5.2}$$

Taking the dual form, similar derivation to Section 4.3.3 - 4.3.5 can be attained with the exception of $g_{\boldsymbol{\delta}}(\boldsymbol{\alpha})$ in problem (4.7) becoming $\frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i(\mathbf{x}_i \odot \boldsymbol{\delta})\right\|^2 + \frac{1}{2C}\boldsymbol{\alpha}'\boldsymbol{\alpha}$, while problem (4.10) takes the form of

$$\min_{\boldsymbol{\mu}\in\mathcal{U}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\Big(\sum_{\boldsymbol{\delta}^t\in\Lambda}\mu^t\mathbf{X}_t\mathbf{X}_t' + \frac{1}{C}\mathbf{I}\Big)\boldsymbol{\alpha}. \tag{5.3}$$

The corresponding feature discriminative score can then be computed with $s_j = \sum_{i=1}^{n}\alpha_i x_{ij}$ $= \sum_{i=1}^{n}\alpha_i f_{ji} = \mathbf{f}_j\boldsymbol{\alpha}$. At the same time, the decision to predict normal pattens can be determined by $f(\mathbf{x}) = \text{sgn}((\mathbf{w}' \odot \boldsymbol{\delta})\mathbf{x} - \gamma)$, where γ is the learned threshold. In this case, Algorithm 2 remains to hold (unless all inputs share the same class), and γ can be obtained from $\boldsymbol{\alpha}$ from the dual solution based on Karush-Kuhn-Tucker (KKT) conditions [184].

The effectiveness of the proposed GDM-OC for solving one class learning problem is showcased here. The standard One-Class SVM (OCSVM) integrated in LIBSVM[1] [185] is then considered for comparison. In the experimental study, all 4 real-world datatsets are considered to maintain consistency. The difference is that only the positive training points of the datasets are used as training data in the one-class setting. Different from binary class setting, since the convergence preference is needed, 0.01% in precision of

[1]http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Table 5.1: Comparison of one class learning result between LIBSVM-OCSVM and GDM.

Dataset	Accuracy		Training Cost (in sec)	
	OCSVM	GDM-OC	OCSVM	GDM-OC
`news20.binary`	**63.11**% ν=0.5	62.8% 56 features	57.77	**1.88**
`kdd2010`	N. A. N. A.	**71.76**% 20 features	30 day +	**973.58**
`webspam`	73.82% ν=0.4	**76.00**% 70 features	7.34×10^5	**1389.9**
`psoriasis`	**52.84**% ν=0.8	47.34% 110 features	2126.15	**388.94**

objective difference is set (In this case, $T = \infty$). $\mathcal{K}_b$ is then set to {2, 5, 10} and a better result is recorded. Further, the ν in LIBSVM-OCSVM is set as {0.2, 0.4, 0.6, 0.8} and the best result is chose for this method that is presented for comparison. The empirical results obtained including accuracy and training cost are given in Table 5.1, from where one can easily figure out that with the embedded feature selection, not only is the training accuracy maintained, the training process is also accelerated. Notably, this speed up is in proportional to the density of the data, as can be observed from Table 4.3. This indicates that GDM suits sparse dataset very well.

5.4 Further Study on The Benefits of Affiliated Features

In this section, some real-world applications, namely digit identification, face recognition and semantic analysis, are studied using various techniques to illustrate the potential benefits of affiliated features and selected feature groups.

5.4.1 Performance of Affiliated Feature Groups on Digit Identification Task

In this subsection, it will be concluded with further details on the interpretation of the proposed GDM algorithm along with the affiliated features attained in Figure 5.5. With

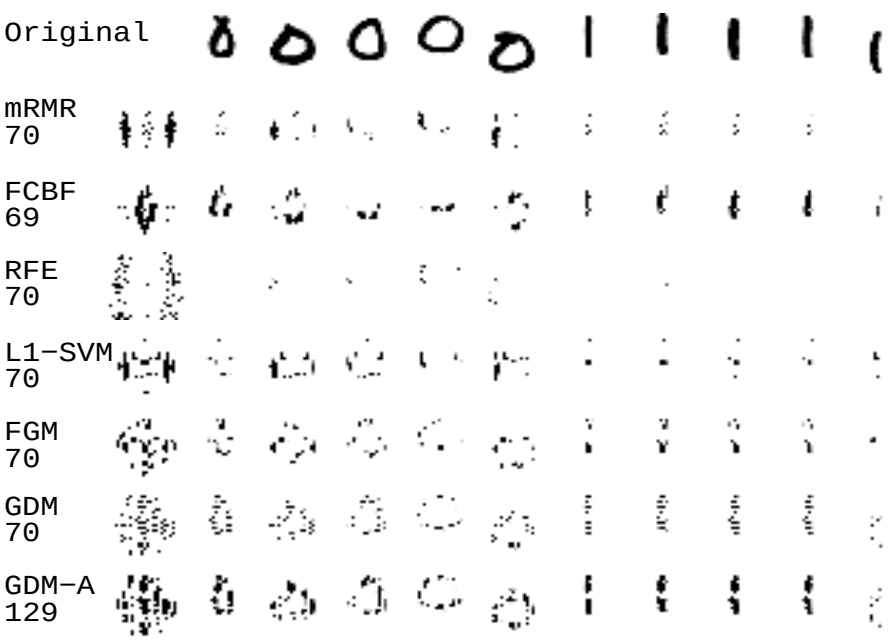


Figure 5.5: Digit identification table results of various methods: example and extracted images by different feature selection methods on `usps` dataset. (Numbers below method indicate the number selected features and the adjacent icons are the overall extracted results. Each element inside the table denotes the superposition result of both digit icon and overall extracted icon.)

respect to the digit identification performance of `usps`[2], the regions highlighted by the affiliated features (129 features) can be considered useful in assisting the human user in identifying a digit "0" or "1" from the extracted images of the original pictures. This is consistent to the observation discussed earlier in Figure 1.2, where the feature groups congregate together in the regions of the beard, mustache and silhouette of the face to form the affiliated feature groups. Hence, information with great significance are reserved for further processing. Referring to the affiliated features in Figure 5.5, despite the highest accuracy achieved by SVM-RFE (the accuracy peak obtained by 70 features), the pixels selected only correspond to the background of the image rather than the black digits, other methods also cannot manifest the clear structure of entire digits befittingly.

[2]Identified digital handwritten characters extracted from the images of "0" and "1" were gathered in the `usps` dataset.

5.4.2 Feature Clustering for Face Recognition

Image clustering is one of the most challenging tasks of computer vision research, especially when dealing with face images. The reason falls on the fact that very often the description of different faces can be rather similar whilst being very distinct only in the background. Thus, traditional clustering methods without proper feature normalization tend to be easily misled. Recently, Nie *et al.* presented a spectral embedded clustering (SEC) method, which learns the feature embedding and clustering at the same time, and reported state-of-the-art face clustering performances on several commonly used face image databases [186]. In this section, to further illustrate the benefits of the proposed GDM, particularly the practicality and interpretability of the derived affiliated features, an intriguing example is showcased where feature groups describe important regions that discriminate between faces and flowers.

The experiment comprises of two core steps: Firstly, the feature groups of face region are identified using a "face vs. non-face" binary scheme. Next, the SEC is conducted on the selected feature groups for further evaluation, in comparison to the original full pixels. The `17 category flower dataset`[3] is adopted for the non-face dataset and the `Yale Face Database B`[4] as the face dataset. The selected features (pixels) or feature groups are shown in Figure 5.6 (*i.e.*, selected important face region), which look like pencil sketches of portraits upon aggregation. Subsequent clustering on the selected features as a multi-class classification task leads to the results given in Table 5.2, where significant improvements in the clustering performance in terms of both clustering accuracy and mutual information are reported. Notably, the running time to perform clustering is also significantly reduced when the feature groups are used over the original pixels as feature descriptors.

5.4.3 Usefulness of Affiliated Features on Text Data

The dataset considered in this subsection is BBC dataset[5], which consists of news articles from the British Broadcasting Corporation. The dataset contains thousands of complete

[3] http://www.robots.ox.ac.uk/~vgg/data/flowers
[4] http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html
[5] http://mlg.ucd.ie/datasets/segment.html

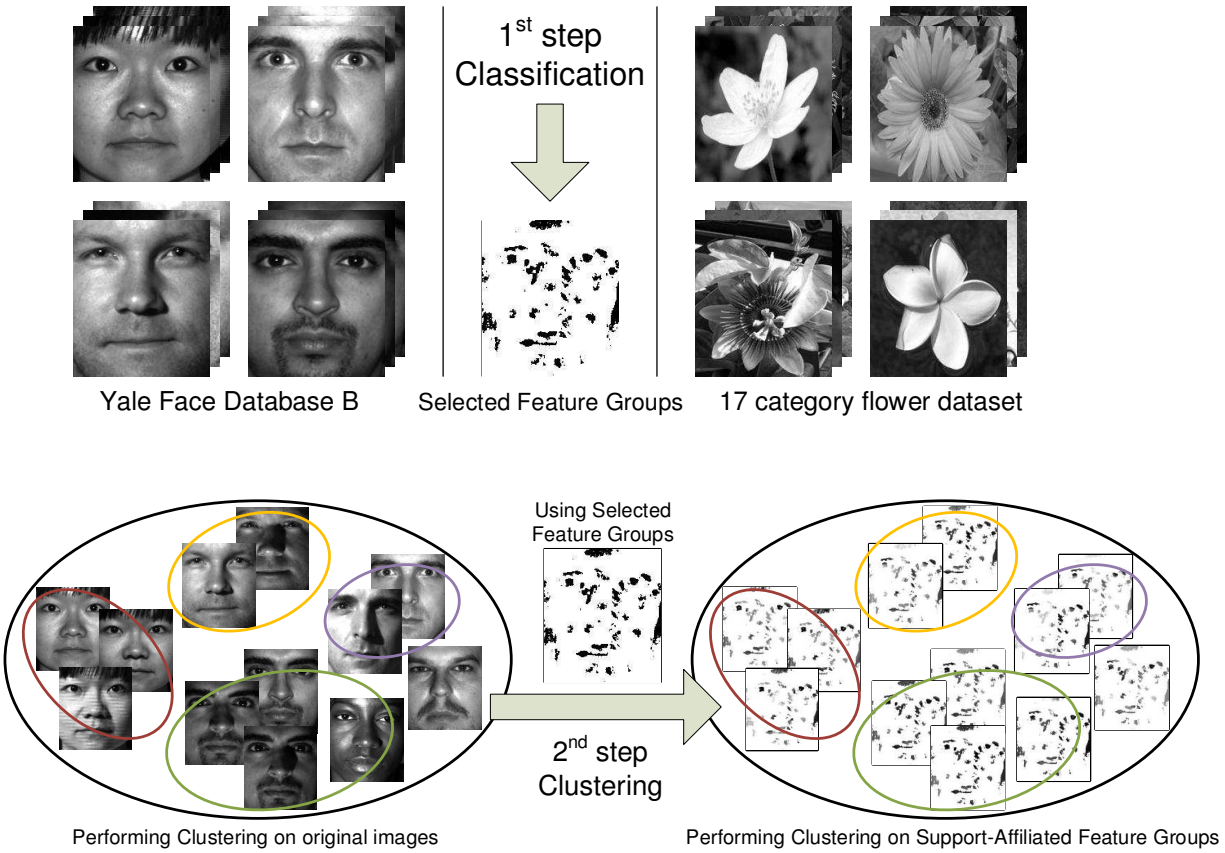


Figure 5.6: Interpretability of selected Support-Affiliated feature groups.

Table 5.2: Clustering performance results using original and selected pixels for face recognition from 5 runs (i.e., the value is in the form of mean±std.).

Features Used	Original	Support-Affiliated Feature Groups
Clustering Accu. (in %)	32.67±0.20	**34.98**±0.36
Mutual Info. (in %)	45.36±0.18	**47.22**±0.29
Elapsed Time (in sec.)	320.15±1.29	**87.49±1.07**

Table 5.3: Details of the BBC dataset.

# Instances						# Features
total	business	entertainment	politics	sport	tech	
1828	414	307	380	351	376	5,470

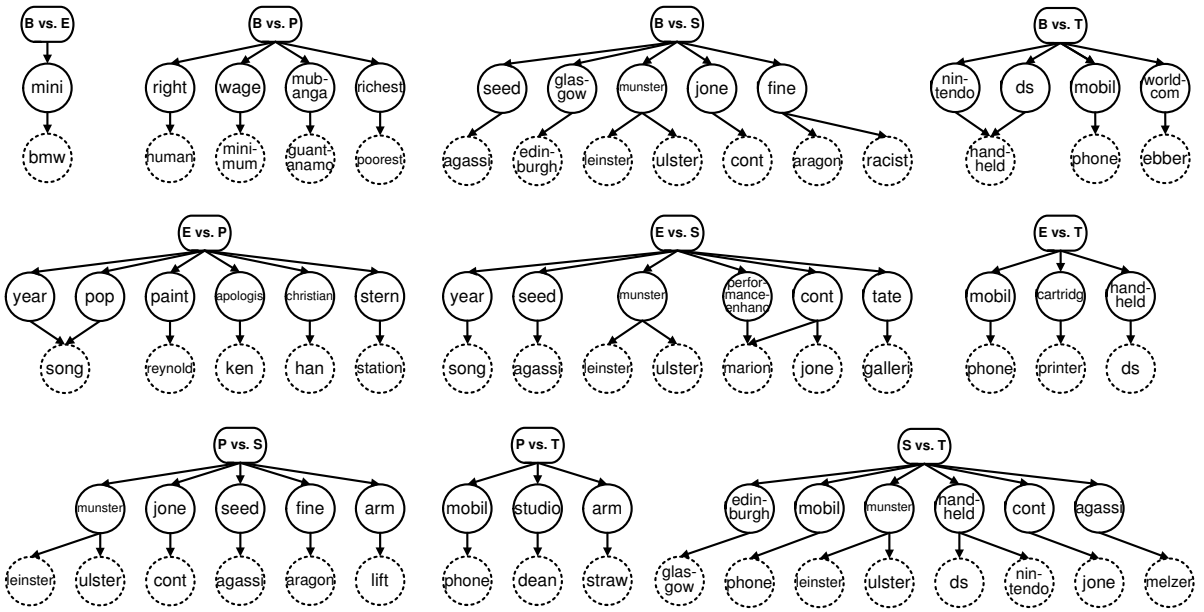


Figure 5.7: Selected support features and affiliated features from the first segment of BBC data.

news articles corresponding to stories in five topical areas including business (B), entertainment (E), politics (P), sport (S) and tech (T), wherein each document is segmented and the segments are randomly assigned. Moreover, for each segmentation, there is a full code-book, indicating the content of each feature. Furthermore, for each feature, the feature value is measured by the frequency of the word. For the experiment, the first segment is picked. The details about this segment is listed in Table 5.3. Correspondingly, the multi-class classification will be achieved based on "one vs. one" strategy, where 10 sets of support-affiliated groups in total will be retrieved from this segment.

The selected results are summarized in Figure 5.7, from where one can discover that many feature groups found are phrases or idiomatic collocations. Similar to the association rule, one can discover the knowledge from some "intuitively irrelevant" words/phrases. For example, in "business vs. politics", there is a feature group of *Mubanga* and *Guantanamo* that discovered by the GDM, where the machine learns the association between these two features, since the original text is: "*According to the Guardian newspaper, Mr.* ***Mubanga*** *alleges torture during his detention at* ***Guantanamo*** *Bay.*"[6]

[6]http://news.bbc.co.uk/1/hi/uk/4163911.stm

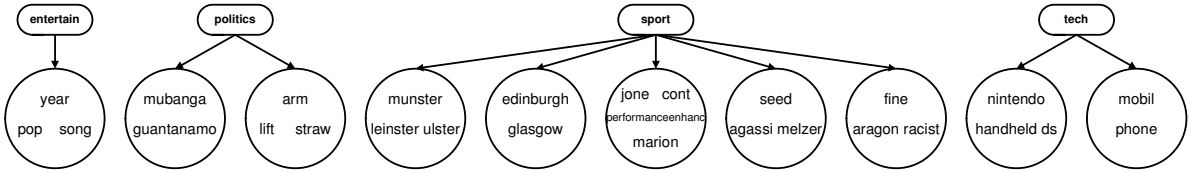


Figure 5.8: Feature groups regarding to each topic.

Furthermore, since this dataset is essentially multi-class based, one can summarize the selected features for each class afterwards. For example, *glasgow* and *edinburgh* have together appeared in both "business vs. sport" and "sport vs. tech", thus this two features could only belong to the class of sport. The summarized results are then drawn in Figure 5.8, wherein the representative features are indicated through the ring, which would further benefit the learning task. On the one hand, one can gather the scattered features to make a more comprehensive understanding of the learning task. For example, in "sport", there is a ring with *seed*, *agassi* and *melzer* from Figure 5.8. And back to the content, we found two pieces of information, namely, "*Agassi is a seed player*" and "*Agassi had a match with Melzer*". On the other hand, since "sport" has the most groups and features, such that this class is easier to be differentiated while "business" news are lack of prominent characteristics in classification. This may further help us to automatically categorize the importance of each class, as well as display the stability of the idiomatic collocations from the news. Moreover, the feature groups selected may help to build accurate classifier from various aspects.

5.5 Conclusion

To summarize, the proposed feature grouping and selection framework, namely GDM, has been shown less sensitive to the key parameters of interest. Besides, the notion and significance of correlated features, namely the support-affiliated feature group has been introduced in the thesis, and how it can benefit the task of feature learning in general since such feature structure can further help to the completion of retrieved knowledge has been also showcased. Further, to identify robust informative features with minimal sampling bias, Embedded Cross Validation is designed to embed the cross validation

scheme in seeking for stable and robust feature sets that are consistent across different data folds. Besides, we also demonstrated the proposed method on One Class Learning, where notable acceleration can be achieved by GDM-OC from big dimensional one-class data. Consequently, the proposed GDM can assist in mining the underlying information from the feature correlations in seconds, offering a new angle to generate more abundant and meaningful results. Hence, exploring novel constraints that are suitable for the discovery of new structures in high dimensional tasks are aspired.

Chapter 6

Concluding Remarks and Future Works

6.1 Concluding Remarks

In this thesis, the notion of "Big Dimensionality" has been firstly introduced. In a similar spirit to "Big Data", the term Big Dimensionality has been coined to put attention on the need for new ways in coping with the unprecedented number of features (dimensions) that are scaling to levels that now renders existing computational intelligence approaches inadequate. Our survey has revealed the lack of studies on the evolution of data dimensionality in the era of Big Data. In particular, our analysis on three popular data repositories has uncovered an exponential increase in the dimensionality of many datasets that have been produced since early 2000s. In life science research, for instance, a quantum leap from the original thousands of genes (features) to millions of Single-Nucleotide Polymorphism in a short period of time has been observed. And there is much evidence that the upward trend of Big Dimensionality will only continue to follow, as influenced by the rapid advancements in computing and information technologies and the arising myriads of feature descriptors, where a forecast of 40 billion features in dimensions is to be expected by year 2020. Based on our detailed analyses, it has been found that the progress of feature selection methods in the field of computational intelligence are falling very much behind the rapidly rising pace of data dimensionality or Big Dimensionality. Last but not least, the core challenges of feature selection (*curse of Big Dimensionality*) and the potential benefits of dimensionality explosion in feature selection (*blessings of Big Dimensionality*) have been presented and discussed.

Since that modern databases with Big Dimensionality are experiencing a growing trend, the state-of-the-art approaches that require the calculations of pairwise feature correlations in their algorithmic designs have not coped well on such database, i.e., the computation burden of m^2 is often impractical. In the latter half of the thesis, the observations from several real-world databases have established that an extremely small portion of the feature pairs contribute significantly to the underlying interactions, i.e., there is a presence of sparse correlations, and there exists feature groups that are highly correlated. Taking the cue, a comprehensive study on potential correlated informative features or feature groups using the concepts of support feature and affiliated feature is presented, to fill in the research gap that has been identified. In particular, the proposed framework, Group Discovery Machine (GDM) embeds an explicit incorporation of both linear and nonlinear correlation measures as constraints in the learning model to filter out large number of non-contributing correlations that could otherwise confuse a classifier, while identifying the correlated and informative feature groups. Notably, the affiliated features are constructed in the proposed method without any additional cost, since they are generated along with the support features. Besides, the proposed method on one class learning is also demonstrated, where notable acceleration can be achieved by GDM-OC from big dimensional one-class data. Further, to identify robust informative features with minimal sampling bias, embedded cross validation is designed to embed the cross validation scheme in seeking for stable and robust feature sets that are consistent across different data folds. Extensive empirical studies have been reported on both synthetic and several real-world datasets to reveal the superior prediction accuracy of GDM through the identified SFs, while some feature redundancies via the identification of AFs are useful for enhanced interpretation of the learning tasks. Last but not least, both the theoretical analysis and empirical studies verified the high efficacies of the proposed methodology, which makes trillion correlations feasible when dealing with big dimensional data.

6.2 Future Works

Through this thesis, it is hoped that this acknowledgement on Big Dimensionality would serve to highlight the need for renewed research efforts in the era of Big Data. Besides,

the methodologies presented in this report have provided some of the groundworks pertaining to the development of feature group structure as well as feature selection on big dimensional data. In what follows, some potential future works are outlined and described.

6.2.1 Memetic Computation

It is worth noting that an interesting line of thought is to use a "divide and conquer" approach to handle this explosion of Big Data and Dimensionality. Specifically, subset of features can be identified from disparate set (subset) of data where memes that have been perceived as building blocks of structured knowledge can be derived via *Deep Learning* [187], *Memetic Computation* [188] or otherwise, for improved scalability, applicability and reusability[1]. These atomized units of memes, metamemes, or memeplex can then be expressed in hierarchical nested relationships or conceptual entities for higher-order learning [190], thus forming societies of the mind for more effective problem solving.

6.2.2 Real Time Data Analytics

With the prevalence of social media networks and portable devices, the demands for sophisticated portable device applications (*e.g.*, video/image concept detection, interest scene detection, spam detection, sentiment analysis, etc.) in handling big **volumes** of multimedia content is rising. In such applications, real-time performance is of utmost importance to users, since no one is willing to spend any time waiting nowadays. In other words, achieving real-time analysis and prediction on these Big Dimensionality is a new challenge of computational intelligence on portable platforms. Due to the potential benefits of this challenging task, this research shall further embark upon such a study as future work.

6.2.3 Transfer Learning for Feature Group Structure

Recently, transfer learning, as a new machine learning strategy, has been successfully employed in many existing machine learning tasks. The essence of this work is to explore

[1]From the statistics generated by Thomson Reuters that targets on world-class emerging research trends, memetic computation has been recently singled out as among the top 10 research fronts in the category of "Mathematics, Computer Science and Engineering" [189].

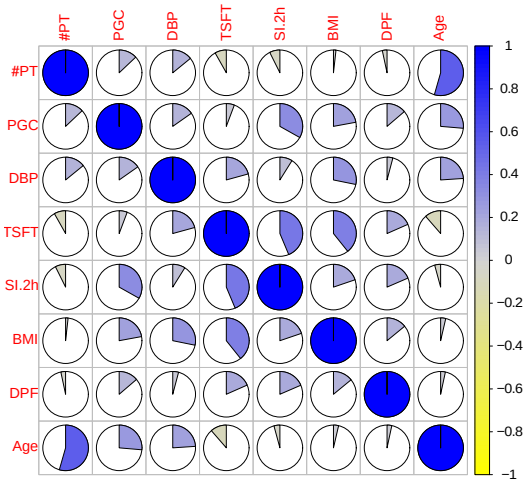
how individuals can transfer from one context towards another context that shared similar characteristics. The research on transfer learning has since attracted much attention in numerous domains, producing a wealth of empirical findings and theoretical interpretations. However, there remains considerable controversy about how transfer learning should be conceptualized and explained. Particularly, the focuses has been on addressing the following questions: what is the probability of occurrence, what is the relation to learning in general, and whether it may be said to exist at all?

Although there have been some existing works on feature transfer learning, such as TrAdaBoost in [191], MMDE in [192], the translated learning model TLRisk in [193], little research has embarked on feature transfer learning involving very high dimensional data nor on the correlation considerations between features. More importantly, the transfer of feature group structures across feature spaces has remained unexplored, wherein seeking the counterparts for both support features and affiliated features would follow the idea.

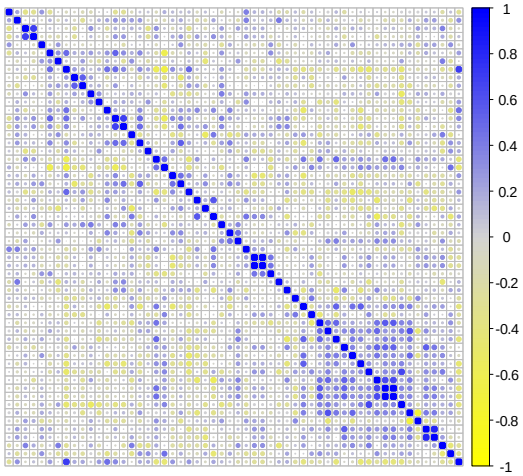
6.2.4 Visualization of Big Dimensional Datasets

To verify and authenticate the identified features, one classical technology that has always been helpful to human experts in data analytics is none other than "visualization". Visual analytics in particularly, has been defined in [194] as *"the science of analytical reasoning facilitated by interactive visual interfaces"*. A presentation of the `diabetes` dataset given in Figure 6.1.a showcases a simple visualization of the correlations between feature pairs using a 2D correlation matrix. With only 8 features in the `diabetes` dataset, the 2D correlation matrix can be embedded with correlation coefficients (using pie-charts) and feature labels (using red fonts) information that would aid in enhancing human verifications and understandings of the data. On the `lung cancer` dataset, which has 7 times features than the `diabetes` dataset, a panoramic view of the 2D correlation matrix as given in Figure 6.1.b remains legible, although the correlation coefficients and feature labels can no longer be included with ease. In the case of the `psoriasis`[2] dataset which comprises 529,651 SNPs (as features or dimensions), a visualization of the 2D correlation matrix involving 280.5 billion grids can hardly make any sense to a human user, referred as 6.1.c. Thus, as the feature size continues to grow and evolve towards the

[2] `http://www.sph.umich.edu/csg/abecasis/casp`



6.1.a: `diabetes` (8 features)



6.1.b: `lung cancer` (56 features)

6.1.c: `psoriasis` (529,651 features)

Figure 6.1: The evolution (rise) of feature dimensionality in correlation matrices.

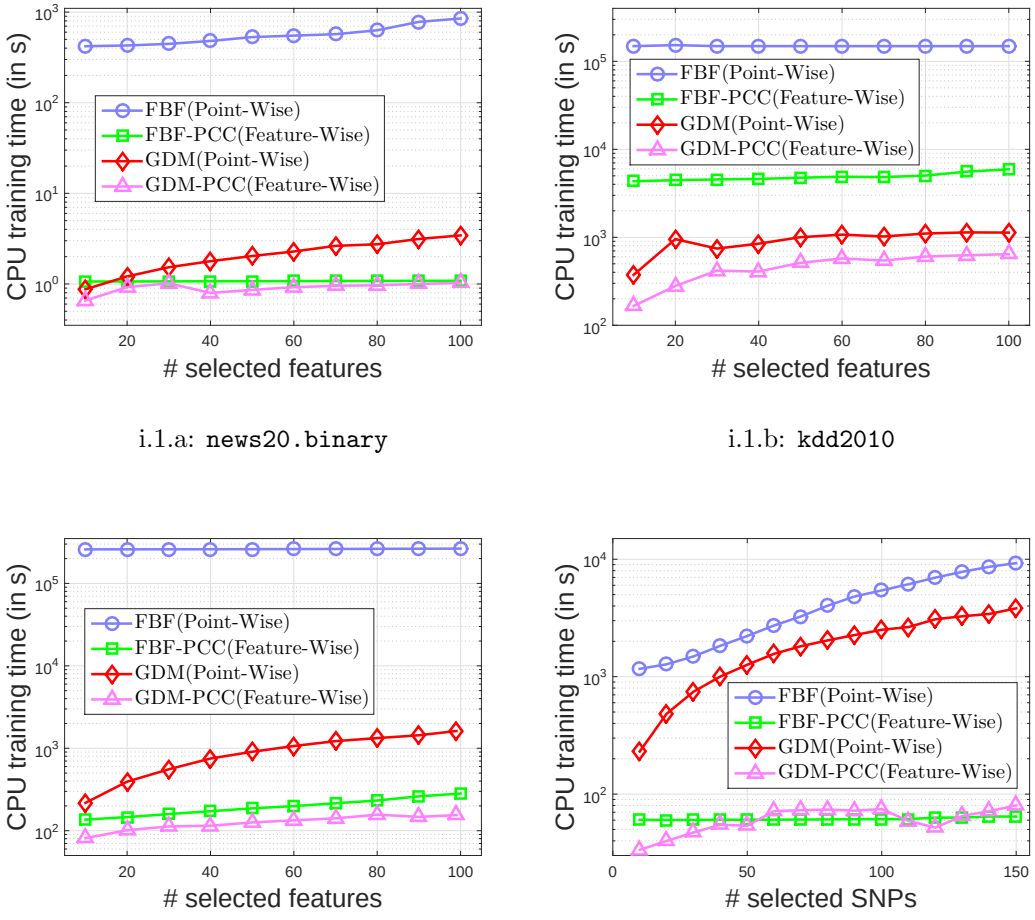
phenomenon of Big Dimensionality, fresh visualization technologies and tools that can equip decision makers with the flexibility to combine creativity and domain knowledge for the identification of features that contain valuable commercial **value**, from the bulk of big dimensional features, would be absolutely essential.

Appendix

i. Experimental Result on Filter-based Brute-force Approach

We noted in the Section 4.3.5 of the main article that "Note that, alternatively, one could employ a brute-force approach to search across all features and pairwise correlations to identify all feature groups that achieves the similar goal. However, such a scheme (i.e., mRMR [161]) can be computationally intensive even with small dataset and would become computational intractable on big dimensional data." There, the brute-force method that we refer to considers a full correlation matrix with $O(m^2)$ complexity. For a mere 10,000-dimension dataset, this translates to 49,995,000 of pairwise computations which can be highly intensive, thus making such a scheme impractical. Note that, besides mRMR, existing state-of-the-art feature grouping methods considered in the manuscript, including OSCAR and GFlasso, consider such a brute-force scheme to prune the correlation/covariance matrix when generating the feature groups.

Notably, the brute-force scheme can be used in place of GDM, for example, by first applying a filter-based method to find the informative features and then simply finding the features that are correlated with the selected features. For the sake of completeness, we have considered such a scheme in our experimental study to assess the simple brute force scheme with the proposed GDM. Specifically, in our experiment, we first apply a correlation measure $\rho(\mathbf{f}_i, \mathbf{y})$ for i^{th} feature to find the most informative features and consequently, using $\rho(\mathbf{f}_i, \mathbf{f}_j)$ for feature grouping. Here we label this method as the Filter-Brute-Force with PCC (FBF-PCC). For the sake of brevity, in the rest of this section, the suffix (i.e., PCC) is omitted and denoted as GDM and FBF, respectively, since only the PCC is considered here.

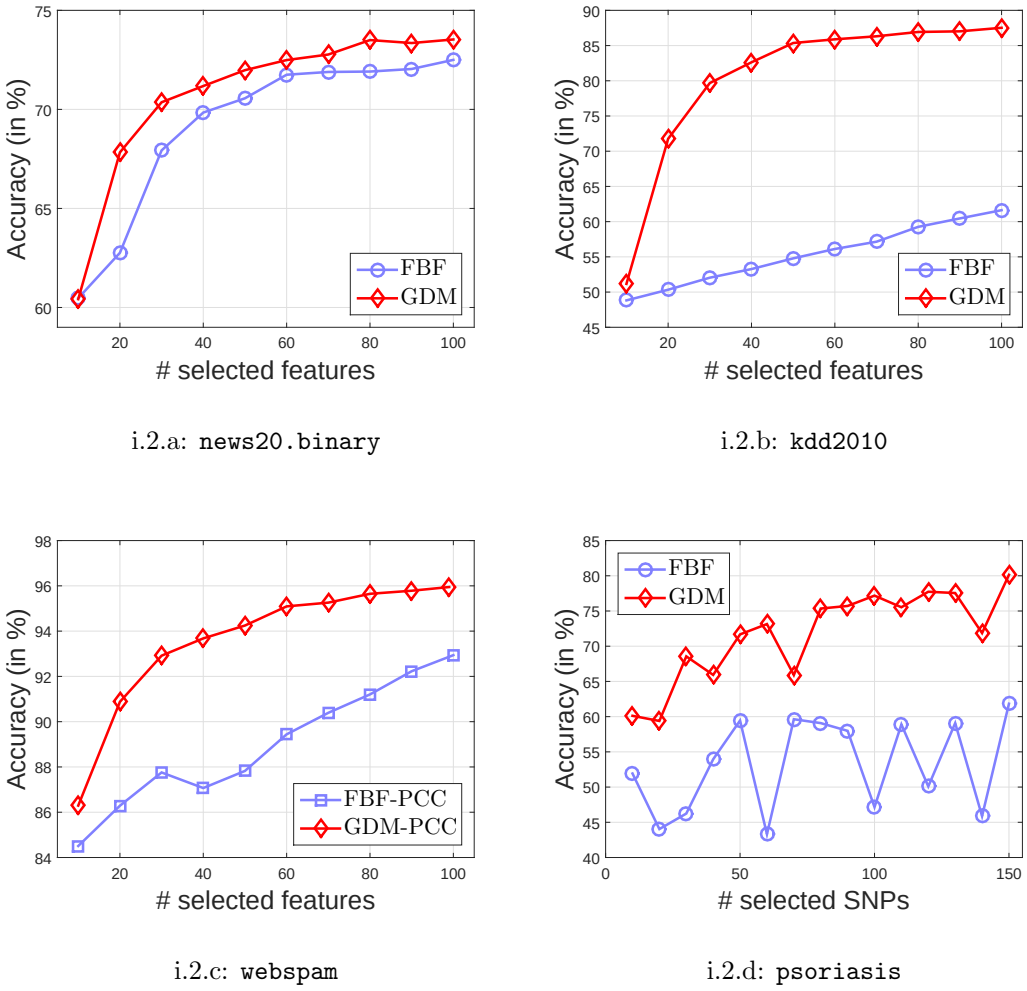


i.1.a: `news20.binary` i.1.b: `kdd2010`

i.1.c: `webspam` i.1.d: `psoriasis`

Figure i.1: A comparison in training time (in seconds) of the GDM and FBF with different implementations on real-world datasets (in logarithmic scale).

From the training time depicted in Figure i.1, one can observe that the FBF incurred a far larger amount of time to compute as compared to the GDM. Noted that this is of discrepancy to the theoretical time complexity derived for such a scheme. It is realized that such discrepancy is due to the point-wise data structure implementation commonly used in many state-of-the-art methods and software packages, such as the LibSVM (i.e., loading each point in a sparse manner; e.g., for each point, [feature index : feature value] is loaded, and wherever the feature value is 0, no memory is allocated). Consequently, when locating a feature, one needs to search from each sample point (i.e., sequential access),



i.2.a: `news20.binary` i.2.b: `kdd2010`

i.2.c: `webspam` i.2.d: `psoriasis`

Figure i.2: Testing accuracy (in %) on real-world datasets.

making the calculations of $\rho(\mathbf{f}_i, \mathbf{y})$ an extreme burden in FBF. Note that, the point-wise version of GDM also suffers from such a problem. Nevertheless, with the blessing of big dimensionality, the number of necessary calculations on pairwise feature correlation is not high for the datasets considered. Hence, the GDM showcased higher efficiency than existing state-of-the-art feature grouping and selection methods in Chapter 4.

Taking the cue, we have proceeded to develop a feature-wise data structure to improve the feature searching process to boost the time incurred by the GDM and FBF. The training time results obtained by the feature-wise methods are also summarized in

Figure i.1, where one can observe the significant speedup by the FBF(Feature-Wise) over FBF(Point-Wise) counterpart and GDM(Feature-Wise) over GDM(Point-Wise). In addition, with a feature-wise data structure, GDM converges efficiently at merely hundreds of seconds on `webspam` (i.e., 23.31 GB of data), while tens of seconds on `psoriasis` (i.e., 11.67 GB of data), even on large-scale dataset `kdd2010` (i.e., 19,264,097 training samples). Further, the GDM(Feature-Wise) also demonstrates faster searching comparing to the FBF(Feature-Wise), especially so on the 3 larger datasets. This can be attributed to the following reasons.

(i) To perform feature selection, the "feature score" is used to rank the feature importance. The feature score of GDM $|s_j|$ (i.e., Equation (10)) is cheaper to compute than that used in FBF[1]. Further, in the formulation of feature score, GDM can take advantage of the sparseness in the dataset. For instance, on the highly sparse dataset `kdd2010` (which has a density of 1.017e-06 as reported in Table 3 of the main manuscript), the GDM reported significant speed up over the FBF.

(ii) The GDM only computes the feature score on support vectors observed in the sparse SVM, while the FBF has to compute all the correlations between features and label. E.g., on `webspam` dataset, an average of 165,990 support vectors have been identified, thus only 59.3% of the sample points are involved in feature score computations. Note that, a significant reduction is achieved in the amount of computations required by the GDM.

Note that, as shown in Figure i.2, GDM also demonstrated improved performance accuracy on all of the datasets considered. In addition, GDM provides a more general framework, since a filter-based brute-force cannot handle specific problems such as the one-class problem. Thus, in a conclusion, although simple filter-based brute-force method can be used for the feature grouping task with simple implementation, it is much less efficient than the GDM.

[1]Traditional filter-based brute force method calculates the dependency between features and label as a criterion to rank the informative level of each feature.

ii. Pseudo Code for Synthetic Data Generation

In the synthetic experiment, the synthetic data is generated to study the capability of a feature grouping method in correctly detecting the ground-truth feature groups set in advance. Algorithm 6 thus outlines the steps of the data generation. Firstly, two random sets $\mathbf{X}$ (for training) and $\bar{\mathbf{X}}$ (for testing) are created with $\mathbb{R}^{9974\times2048}$ insize. Consequently, 12 features are configured as informative features, while inducing 26 predefined[2] affiliated features to arrive at a dataset comprising 10,000 dimensions (i.e., $\mathbb{R}^{(9974+26)\times2048}$).

Predefine 12 SFs and corresponding AFs (26 in total)
Set $n = 2048$ and $m = 9974$
$\mathbf{X} = \text{randn}(m, n)$ (for training data)
$\bar{\mathbf{X}} = \text{randn}(m, n)$ (for testing data)
Set all the indices of the SFs and AFs to an index set $\mathcal{Q}$.
for $i = 1$ **to** 12 (size of SFs) **do**
  Perform linear transformation to generate AFs (26 features in total) for the SF w.r.t. the size defined in Figure 4.4.a.
end for
1. Add the AFs into $\mathbf{X}$ and $\bar{\mathbf{X}}$ while accommodating the indices, leading $\mathbf{X}$ and $\bar{\mathbf{X}}$ towards 10,000 in dimension
2. Set a vector $\mathbf{w}$ with $\mathbf{w}$ = zeros(m,1)
3. $\mathbf{w}(\mathcal{Q})$ = rand(length($\mathcal{Q}$),1) %generate weight value only for informative features (SFs + AFs)
4. $\mathbf{y} = \mathbf{w}'\mathbf{X}$ and $\bar{\mathbf{y}} = \mathbf{w}'\bar{\mathbf{X}}$ %generate label
5. Set labels as $\mathbf{y}(\mathbf{y}\geqslant 0) = 1$ $\mathbf{y}(\mathbf{y}<0) = -1$, same to $\bar{\mathbf{y}}$

Algorithm 6: Synthetic Data Generator in MATLAB.

In addition, the affiliated features of the dataset are generated by making linear superposition, perturbation, as well as angle transformation to the corresponding support feature, leading to linear correlated feature groups. Note that, one can also adapt non-linear correlations in data generation for GDM-SU.

[2]Here the word "predefined" indicates that only the indices of the features are defined in advance. The weight vector $\mathbf{w}$ is generated to make sure that the predefined features are informative and the labels are then set by the output of $\mathbf{w}'\mathbf{X}$.

References

[1] Bing-Yu Sun, Xiaoming Zhang, Jiuyong Li, and Xue-Min Mao. Feature Fusion Using Locally Linear Embedding for Classification. *IEEE Trans. Neural Netw.*, 21(1):163–168, 2010.

[2] Junbin Gao, Jun Zhang, and David Tien. Relevance Units Latent Variable Model and Nonlinear Dimensionality Reduction. *IEEE Trans. Neural Netw.*, 21(1):123–135, 2010.

[3] Sang Wan Lee and Zeungnam Bien. Representation of a Fisher Criterion Function in a Kernel Feature Space. *IEEE Trans. Neural Netw.*, 21(2):333–339, 2010.

[4] Pietari Pulkkinen and Hannu Koivisto. A Dynamically Constrained Multiobjective Genetic Fuzzy System for Regression Problems. *IEEE Trans. Fuzzy Syst.*, 18(1):161–177, 2010.

[5] Yaman Aksu, David J. Miller, George Kesidis, and Qing X. Yang. Margin-Maximizing Feature Elimination Methods for Linear and Nonlinear Kernel-Based Discriminant Functions. *IEEE Trans. Neural Netw.*, 5(21):710–717, May 2010.

[6] Zexuan Zhu, Sen Jia, and Zhen Ji. Towards a Memetic Feature Selection Paradigm. *IEEE Comput. Intell. Mag.*, 5(2):41–53, May 2010.

[7] Luping Zhou, Lei Wang, and Chunhua Shen. Feature Selection With Redundancy-Constrained Class Separability. *IEEE Trans. Neural Netw.*, 21(5):853–858, 2010.

[8] Dudy Lim, Yaochu Jin, Yew-Soon Ong, and Bernhard Sendhoff. Generalizing Surrogate-assisted Evolutionary Computation. *IEEE Trans. Evol. Comput.*, 14(3):329–355, May 2010.

[9] María José Gacto, Rafael Alcalá, and Francisco Herrera. Integration of an Index to Preserve the Semantic Interpretability in the Multiobjective Evolutionary Rule Selection and Tuning of Linguistic Fuzzy Systems. *IEEE Trans. Fuzzy Syst.*, 18(3):515–531, 2010.

[10] Kai Zhang and James T. Kwok. Clustered Nyström Method for Large Scale Manifold Learning and Dimension Reduction. *IEEE Trans. Neural Netw.*, 21(10):1576–1587, 2010.

[11] Hua Huang and Huiting He. Super-Resolution Method for Face Recognition Using Nonlinear Mappings on Coherent Features. *IEEE Trans. Neural Netw.*, 22(1):121–130, 2011.

[12] Michael G. Epitropakis, Dimitris K. Tasoulis, Nicos G. Pavlidis, Vassilis P. Plagianakos, and Michael N. Vrahatis. Enhancing Differential Evolution Utilizing Proximity-Based Mutation Operators. *IEEE Trans. Evol. Comput.*, 15(1):99–119, 2011.

[13] Xiaowei Yang, Guangquan Zhang, Jie Lu, and Jun Ma. A Kernel Fuzzy c-Means Clustering-Based Fuzzy Support Vector Machine Algorithm for Classification Problems With Outliers or Noises. *IEEE Trans. Fuzzy Syst.*, 19(1):105–115, 2011.

[14] Cheng-Hsuan Li, Bor-Chen Kuo, and Chin-Teng Lin. LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction. *IEEE Trans. Fuzzy Syst.*, 19(1):152–163, 2011.

[15] Zhaohong Deng, Kup-Sze Choi, Fu-Lai Chung, and Shitong Wang. Scalable TSK Fuzzy Modeling for Very Large Datasets Using Minimal-Enclosing-Ball Approximation. *IEEE Trans. Fuzzy Syst.*, 19(2):210–226, 2011.

[16] Rami N. Mahdi and Eric C. Rouchka. Reduced HyperBF Networks: Regularization by Explicit Complexity Reduction and Scaled Rprop-Based Training. *IEEE Trans. Neural Netw.*, 22(5):673–686, 2011.

[17] Jian-Bo Yang and Chong Jin Ong. Feature Selection Using Probabilistic Prediction of Support Vector Regression. *IEEE Trans. Neural Netw.*, 22(6):954–962, 2011.

[18] Hemant Kumar Singh, Amitay Isaacs, and Tapabrata Ray. A Pareto Corner Search Evolutionary Algorithm and Dimensionality Reduction in Many-Objective Optimization Problems. *IEEE Trans. Evol. Comput.*, 15(4):539–556, 2011.

[19] Rafael Alcalá, María José Gacto, and Francisco Herrera. A Fast and Scalable Multiobjective Genetic Fuzzy System for Linguistic Fuzzy Modeling in High-Dimensional Regression Problems. *IEEE Trans. Fuzzy Syst.*, 19(4):666–681, 2011.

[20] Bo Wang, Shuming Wang, and Junzo Watada. Fuzzy-Portfolio-Selection Models With Value-at-Risk. *IEEE Trans. Fuzzy Syst.*, 19(4):758–769, 2011.

[21] Jesús Alcalá-Fdez, Rafael Alcalá, and Francisco Herrera. A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning. *IEEE Trans. Fuzzy Syst.*, 19(5):857–872, 2011.

[22] Farid Oveisi, Shahrzad Oveisi, Abbas Erfanian, and Ioannis Patras. Tree-Structured Feature Extraction Using Mutual Information. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(1):127–137, 2012.

[23] Qinghua Hu, Weiwei Pan, Lei Zhang, David Zhang, Yanping Song, Maozu Guo, and Daren Yu. Feature Selection for Monotonic Classification. *IEEE Trans. Fuzzy Syst.*, 20(1):69–81, 2012.

[24] Yi Huang, Dong Xu, and Feiping Nie. Semi-Supervised Dimension Reduction Using Trace Ratio Criterion. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(3):519–526, 2012.

[25] André Stuhlsatz, Jens Lippel, and Thomas Zielke. Feature Extraction With Deep Neural Networks by a Generalized Discriminant Analysis. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(4):596–608, 2012.

[26] Xiaodong Li and Xin Yao. Cooperatively Coevolving Particle Swarms for Large Scale Optimization. *IEEE Trans. Evol. Comput.*, 16(2):210–224, 2012.

[27] Michela Antonelli, Pietro Ducange, and Francesco Marcelloni. Genetic Training Instance Selection in Multiobjective Evolutionary Fuzzy Systems: A Coevolutionary Approach. *IEEE Trans. Fuzzy Syst.*, 20(2):276–290, 2012.

[28] Yuxi Hou, Iickho Song, Hwang-Ki Min, and Cheol Hoon Park. Complexity-Reduced Scheme for Feature Extraction With Linear Discriminant Analysis. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(6):1003–1009, 2012.

[29] Edmund K. Burke, Matthew R. Hyde, and Graham Kendall. Grammatical Evolution of Local Search Heuristics. *IEEE Trans. Evol. Comput.*, 16(3):406–417, 2012.

[30] Tim Blackwell. A Study of Collapse in Bare Bones Particle Swarm Optimization. *IEEE Trans. Evol. Comput.*, 16(3):354–372, 2012.

[31] Chun-Wei Seah, Ivor W. Tsang, and Yew-Soon Ong. Transductive Ordinal Regression. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(7):1074–1086, 2012.

[32] Yi-Cheng Chen, Nikhil R. Pal, and I-Fang Chung. An Integrated Mechanism for Feature Selection and Fuzzy Rule Extraction for Classification. *IEEE Trans. Fuzzy Syst.*, 20(4):683–698, 2012.

[33] Farhad Hassanzadeh, Mikael Collan, and Mohammad Modarres. A Practical Approach to R&D Portfolio Selection Using the Fuzzy Pay-Off Method. *IEEE Trans. Fuzzy Syst.*, 20(4):615–622, 2012.

[34] Leon Wenliang Zhong and James T. Kwok. Efficient Sparse Modeling With Automatic Feature Grouping. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(9):1436–1447, 2012.

[35] Praisan Padungweang, Chidchanok Lursinsap, and Khamron Sunat. A Discrimination Analysis for Unsupervised Feature Selection via Optic Diffraction Principle. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(10):1587–1600, 2012.

[36] Mathieu Ramona, Gaël Richard, and Bertrand David. Multiclass Feature Selection With Kernel Gram-Matrix-Based Criteria. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(10):1611–1623, 2012.

[37] Kourosh Neshatian, Mengjie Zhang, and Peter Andreae. A Filter Approach to Multiple Feature Construction for Symbolic Learning Classifiers Using Genetic Programming. *IEEE Trans. Evol. Comput.*, 16(5):645–661, 2012.

[38] Shiming Xiang, Feiping Nie, Gaofeng Meng, Chunhong Pan, and Changshui Zhang. Discriminative Least Squares Regression for Multiclass Classification and Feature Selection. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(11):1738–1754, 2012.

[39] Timothy Hancock and Hiroshi Mamitsuka. Boosted Network Classifiers for Local Feature Selection. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(11):1767–1778, 2012.

[40] Timothy C. Havens, James C. Bezdek, Christopher Leckie, Lawrence O. Hall, and Marimuthu Palaniswami. Fuzzy c-Means Algorithms for Very Large Data. *IEEE Trans. Fuzzy Syst.*, 20(6):1130–1146, 2012.

[41] Grzegorz Dudek. An Artificial Immune System for Classification With Local Feature Selection. *IEEE Trans. Evol. Comput.*, 16(6):847–860, 2012.

[42] Jing Chen, Zhengming Ma, and Yang Liu. Local Coordinates Alignment With Global Preservation for Dimensionality Reduction. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(1):106–117, 2013.

[43] Qi Mao and Ivor W. Tsang. Efficient Multi-Template Learning for Structured Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 24(2):248–261, 2013.

[44] Dhish Kumar Saxena, João A. Duro, Ashutosh Tiwari, Kalyanmoy Deb, and Qingfu Zhang. Objective Reduction in Many-Objective Optimization: Linear and Nonlinear Algorithms. *IEEE Trans. Evol. Comput.*, 17(1):77–99, 2013.

[45] Fernando E. B. Otero, Alex Alves Freitas, and Colin G. Johnson. A New Sequential Covering Strategy for Inducing Classification Rules With Ant Colony Algorithms. *IEEE Trans. Evol. Comput.*, 17(1):64–76, 2013.

[46] Huayang Xie and Mengjie Zhang. Parent Selection Pressure Auto-Tuning for Tournament Selection in Genetic Programming. *IEEE Trans. Evol. Comput.*, 17(1):1–19, 2013.

[47] Wendy Ashlock and Suprakash Datta. Evolved Features for DNA Sequence Classification and Their Fitness Landscapes. *IEEE Trans. Evol. Comput.*, 17(2):185–197, 2013.

[48] Ata Kaban. Fractional Norm Regularization: Learning With Very Few Relevant Features. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(6):953–963, 2013.

[49] José Antonio Sanz, Alberto Fernández, Humberto Bustince, and Francisco Herrera. IVTURS: A Linguistic Fuzzy Rule-Based Classification System Based On a New Interval-Valued Fuzzy Reasoning Method With Tuning and Rule Selection. *IEEE Trans. Fuzzy Syst.*, 21(3):399–411, 2013.

[50] Urvesh Bhowan, Mark Johnston, Mengjie Zhang, and Xin Yao. Evolving Diverse Ensembles Using Genetic Programming for Classification With Unbalanced Data. *IEEE Trans. Evol. Comput.*, 17(3):368–386, 2013.

[51] Edwin Lughofer and Oliver Buchtala. Reliable All-Pairs Evolving Fuzzy Classifiers. *IEEE Trans. Fuzzy Syst.*, 21(4):625–641, 2013.

[52] Mingkui Tan, Ivor W. Tsang, and Li Wang. Minimax Sparse Logistic Regression for Very High-Dimensional Feature Selection. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(10):1609–1622, 2013.

[53] Hesam Izakian, Witold Pedrycz, and Iqbal Jamal. Clustering Spatiotemporal Data: An Augmented Fuzzy C-Means. *IEEE Trans. Fuzzy Syst.*, 21(5):855–868, 2013.

[54] Aniruddha Basak, Swagatam Das, and Kay Chen Tan. Multimodal Optimization Using a Biobjective Differential Evolution Algorithm Enhanced With Mean Distance-Based Selection. *IEEE Trans. Evol. Comput.*, 17(5):666–685, 2013.

[55] Grigorios Skolidis and Guido Sanguinetti. Semisupervised Multitask Learning With Gaussian Processes. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(12):2101–2112, 2013.

[56] Kartick Subramanian, Sundaram Suresh, and Narasimhan Sundararajan. A Metacognitive Neuro-Fuzzy Inference System (McFIS) for Sequential Classification Problems. *IEEE Trans. Fuzzy Syst.*, 21(6):1080–1095, 2013.

[57] Weishan Dong, Tianshi Chen, Peter Tino, and Xin Yao. Scaling Up Estimation of Distribution Algorithms for Continuous Optimization. *IEEE Trans. Evol. Comput.*, 17(6):797–822, 2013.

[58] Hiroshi Someya. Striking a Mean- and Parent-Centric Balance in Real-Valued Crossover Operators. *IEEE Trans. Evol. Comput.*, 17(6):737–754, 2013.

[59] Yiteng Zhai, Yew-Soon Ong, and Ivor W. Tsang. The Emerging 'Big Dimensionality'. *IEEE Comput. Intell. Mag.*, 9(3):14–26, August 2014.

[60] Ping Li et al. Hashing algorithms for large-scale learning. In *Advances in neural information processing systems*, 2011.

[61] Wei Fan and Albert Bifet. Mining big data: current status, and forecast to the future. In *ACM SIGKDD Explorations Newsletter*, volume 14, pages 1–5, 2013.

[62] Daniel E. O'Leary. Artificial Intelligence and Big Data. *IEEE Intelligent Systems*, 28(2):96–99, 2013.

[63] Mike Ferguson. Architecting A Big Data Platform for Analytics. *A Whitepaper Prepared for IBM*, 2012.

[64] John Gantz and David Reinsel. Extracting Value from Chaos, June 2011.

[65] Peer Kröger. Going Big in Data Dimensionality - Challenges and Solutions in Mining High Dimensional Data. In *Symposium on Scalable Analytics*, Garching bei München, November 2012.

[66] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Identifying suspicious urls: An application of large-scale online learning. In *Proceedings of the 26st International Conference on Machine Learning (ICML 2009)*, pages 681–688, Montreal, Canada, June 2009.

[67] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Beyond blacklists: Learning to detect malicious web sites from suspicious urls. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009)*, pages 1245–1253, Paris, France, June 2009.

[68] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[69] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding: a general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:40–51, January 2007.

[70] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

[71] Dell Zhang and Wee Sun Lee. Extracting key-substring-group features for text classification. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 474–483, Philadelphia, Pennsylvania, United States, August 2006.

[72] Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski, and Michael W. Mahoney. Feature selection methods for text classification. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 230–239, San Jose, California, United States, August 2007.

[73] Ben Blum, Michael I. Jordan, David Kim, Rhiju Das, Philip Bradley, and David Baker. Feature selection methods for improving protein structure prediction with rosetta. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS 2007)*, Vancouver, B.C., Canada, December 2007. MIT Press.

[74] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[75] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. Unsupervised feature selection for the k-means clustering problem. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009)*, pages 153–161, Vancouver, B.C., Canada, December 2009.

[76] Seyoung Kim and Eric P. Xing. Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *PLoS Genet.*, 5(8):e1000587, 2009.

[77] Xiaolin Yang, Seyoung Kim, and Eric P. Xing. Heterogeneous Multitask Learning with Joint Sparsity Constraints. In *NIPS*, Vancouver, B.C., Canada, 2009.

[78] M. A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999.

[79] I. Guyon. *Practical Feature Selection: from Correlation to Causality*. IOS Press, 2008.

[80] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu. Advancing feature selection research. Technical report, Arizona State University, 2011.

[81] Z. Zhao et al. On similarity preserving feature selection. *IEEE Trans. Knowl. Data Eng.*, 25(3):619–632, 2013.

[82] J. O'Sullivan, J. Langford, R. Caruana, and A. Blum. Featureboost: A meta learning algorithm that improves model robustness. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 703–710, Stanford, California, United States, 2000.

[83] R. Caruana and V. R. de Sa. Benefitting from the variables that variable selection discards. *J. Mach. Learn. Res.*, 3:1245 – 1264, 2003.

[84] H. Xu, S. Mannor, and C. Caramanis. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):187–193, 2012.

[85] Lei Han and Yu Zhang. Discriminative feature grouping. In *AAAI*, Austin, Texas, United States, 2015.

[86] Leon Wenliang Zhong and James T. Kwok. Efficient sparse modeling with automatic feature grouping. In *ICML*, Bellevue, WA, USA, 2011.

[87] IBM Big Data Success Stories. Technical report, IBM, USA, October 2011.

[88] IBM Big Data Platform - Bringing Big Data to The Enterprise.

[89] Paul C. Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, and James Giles. *Harness the Power of Big Data The IBM Big Data Platform*. McGraw-Hill, 2012.

[90] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H. Byers. Big Data: The Next Frontier for Innovation, Competition and Productivity. Technical report, McKinsey Global Institute, May 2011.

[91] Ashutosh Garg, Sariel Har-Peled, and Dan Roth. On Generalization Bounds, Projection Profile, and Margin Distribution. In *ICML*, pages 171–178, Sydney, Australia, 2002.

[92] Robert J. Durrant and Ata Kabán. Sharp Generalization Error Bounds for Randomly-projected Classifiers. In *ICML*, pages 693–701, Atlanta, USA, 2013.

[93] Peter J. Bickel and Elizaveta Levina. Some Theory for Fishers Linear Discriminant Function, Naive Bayes, and Some Alternatives When There Are Many More Variables Than Observations. *Bernoulli*, 10(6):989–1010, December 2004.

[94] Emmanuel Candes and Terence Tao. The Dantzig Selector: Statistical Estimation When p Is Much Larger than n. *The Annals of Statistics*, 35(6):2313–2351, December 2007.

[95] Peter Gehler and Sebastian Nowozin. On Feature Combination for Multiclass Object Classification. In *ICCV*, Kyoto, Japan, 2009.

[96] The 1st International Workshop on High Dimensional Data Mining (HDM), December.

[97] Special Session for WCCI 2014 "EC Generalisation in High-dimensional Input Spaces".

[98] Special Session for WCCI 2016 "Evolutionary Feature Selection and Construction".

[99] Jin Huang and Charles X. Ling. Constructing New and Better Evaluation Measures for Machine Learning. In *IJCAI*, Hyderabad, India, 2007.

[100] Qi Mao and Ivor W. Tsang. A Feature Selection Method for Multivariate Performance Measures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2051–2063, 2013.

[101] Hugh Leather, Edwin Bonilla, and Michael O'Boyle. Automatic Feature Generation for Machine Learning Based Optimizing Compilation. In *CGO*, pages 81–91, Seattle, WA, USA, March 2009.

[102] Yu Sun and Bir Bhanu. Image Retrieval with Feature Selection and Relevance Feedback. In *ICIP*, pages 3209–3212, Hong Kong, China, 2010.

[103] Maeve Duggan and Lee Rainie. Cell Phone Activities 2012. Technical report, Washington, D.C., November 2012.

[104] Wade Roush. TR10: Peering into Video's Future. *MIT Technology Review March 12, 2007*, March.

[105] YouTube Statistic, 2013.

[106] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[107] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893, San Diego, CA, USA, 2005.

[108] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Visions*, 60(2):91–110, November 2004.

[109] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Journal Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[110] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[111] Andrea Vedaldi and Stefano Soatto. Quick Shift and Kernel Methods for Mode Seeking. In *ECCV*, pages 705–718, Marseille, France, 2008.

[112] James L. Crowley and Alice C. Parker. A Representation for Shape Based on Peaks and Ridges in the Difference of Low Pass Transform. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(2):156–170, March 1984.

[113] Ivan Laptev, Marcin Marszaλek, Cordelia Schmid, and Benjamin Rozenfeld. Learning Realistic Human Actions from Movies. In *CVPR*, Anchorage, AK, USA, 2008.

[114] George Toderici, Hrishikesh Aradhye, Marius Pasca, Luciano Sbaiz, and Jay Yagnik. Finding Meaning on YouTube: Tag Recommendation and Category Discovery. In *CVPR*, 2010.

[115] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In *CVPR*, pages 3169–3176, Colorado Springs, USA, June 2011.

[116] Zhongwen Xu, Yi Yang, Ivor W. Tsang, Nicu Sebe, and Alexander Hauptmann. Feature Weighting via Optimal Thresholding for Video Analysis. In *ICCV*, pages 3440–3447, Sydney, Australia, December 2013.

[117] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics*, 7(3), 2006.

[118] I. Guyon, J. Weston, M.D. S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46:389–422, 2002.

[119] K. Bache and M. Lichman. UCI Machine Learning Repository, 2013.

[120] S. Hettich and S. D. Bay. The UCI KDD Archive, 1999.

[121] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3)(27), 2011.

[122] S. Orrù, E. Giuressi, M. Casula, A. Loizedda, R. Murru, M. Mulargia, M.V. Masala, D. Cerimele, M. Zucca, N. Aste, P. Biggio, C. Carcassi, and L. Contu. Psoriasis is associated with a SNP haplotype of the corneodesmosin gene (CDSN). *Tissue Antigens*, 60(4):292–298, 2002.

[123] Anthony J. Brookes. The essence of SNPs. *Gene*, 234(2):177–186, July 1999.

[124] David L. Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. In *American Mathematical Society Conf. Math Challenges of the 21st Century*. 2000.

[125] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[126] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2nd edition, 2003.

[127] Sen Yang, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. Feature Grouping and Selection Over an Undirected Graph. In *KDD*, Beijing, China, August 2012.

[128] Howard D. Bondell and Brian J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64:115–123, March 2008.

[129] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, November 2006.

[130] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[131] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[132] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[133] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, pages 856–863, Washington, D.C., United States, 2003.

[134] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[135] Z. Zhu, Y.-S. Ong, and M. Dash. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40(11):3236–3248, 2007.

[136] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An improved glmnet for l1-regularized logistic regression and support vector machines. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011)*, pages 33–41, San diego, California, United States, 2011. ACM Press.

[137] M. Tan, L. Wang, and I. W. Tsang. Learning sparse svm for feature selection on very high dimensional datasets. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 1047–1054, Haifa, Israel, 2010.

[138] Q. Mao and I. W. Tsang. Optimizing performance measures for feature selection. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2011)*, pages 1170–1175, Vancouver, Canada, 2011.

[139] Huiqing Liu, Jinyan Li, and Limsoon Wong. A Comparative Study on Feature Selection and Classi cation Methods Using Gene Expression Pro les and Proteomic Patterns. *Genome Informatics*, 13:51–60, 2002.

[140] Chris Ding. Analysis of gene expression profiles: class discovery and leaf ordering. In *Proceedings of the 6th International Conference on Research in Computational Molecular Biology (RECOMB 2002)*, pages 127–136, Washington, D.C., United States, April 2002.

[141] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan. Trace ratio criterion for feature selection. In *Proceedings of the 23rd Association for the Advancement of Artificial Intelligence (AAAI 2008)*, volume 2, pages 671–676, Chicago, Illinois, United States, 2008. AAAI Press.

[142] William H. Press et al. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, 2nd edition, February 1993.

[143] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, July 1994.

[144] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, March 2012.

[145] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Proceedings of the 13th International Conference on Machine Learning (ICML 1996)*, pages 284–292, Bari, Italy, 1996. Morgan Kaufmann.

[146] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, December 1997.

[147] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, September 1998.

[148] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[149] Alain Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, March 2003.

[150] P.S. Bradley and O. L. Mangasarian. Feature Selection via Concave Minimization and Support Vector Machines. In *ICML*, pages 82–90, Madison, Wisconsin, United States, 1998. Morgan Kaufmann.

[151] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm Support Vector Machines. In *NIPS*, Vancouver, B.C., Canada, 2003.

[152] Glenn M. Fung and O. L. Mangasarian. A Feature Selection Newton Method for Support Vector Machine Classification. *Computational Optimization and Applications*, 28(2):185–202, July 2004.

[153] Mário A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, December 2007.

[154] Jr. J. E. Kelley. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[155] A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.

[156] H. Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:433–459, 2010.

[157] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[158] Chris Ding, Xiaofeng He, Hongyuan Zha, and Horst D. Simon. Adaptive dimension reduction for clustering high dimensional data. In *Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM 2002)*, pages 147–154, Maebashi, Japan, December 2002.

[159] Achmad Widodo and Bo-Suk Yang. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Expert Systems with Applications*, 33(1):241–250, July 2007.

[160] Eyal Krupka, Amir Navot, and Naftali Tishby. Learning to select features using their properties. *Journal of Machine Learning Research*, 9:2349–2376, October 2008.

[161] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

[162] L. Yu and H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *ICML*, 2003.

[163] Yiteng Zhai, M. Tan, I. W. Tsang, and Y.-S. Ong. Discovering Support and Affiliated Features from Very High Dimensions. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1455–1462, Edinburgh, Scotland, United Kingdom, July 2012. Omnipress.

[164] R. Paturi, S. Rajasekaran, and J. Reif. The light bulb problem. *Information and Computation*, 117(2):187–192, 1995.

[165] Panagiotis Achlioptas, Bernhard Schölkopf, and Karsten Borgwardt. Two-locus association mapping in subquadratic time. In *KDD*, pages 726–734, 2011.

[166] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *FOCS*, pages 11–20, 2012.

[167] Karl Pearson. *The Life, Letters and Labours of Francis Galton (3 vols. in 4 parts).* Cambridge Univ. Press, 1914-1930.

[168] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[169] Usama M. Fayyad and Keki B. lrani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993)*, pages 1022–1027, Chambery, France, 1993.

[170] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 217–226, Philadelphia, Pennsylvania, United States, August 2006.

[171] Yu-Feng Li, Ivor W. Tsang, James T. Kwok, and Zhi-Hua Zhou. Tighter and convex maximum margin clustering. In *AISTATS*, volume 5, pages 344–351, Clearwater Beach, Florida, USA, April 2009.

[172] Gert R.G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, January 2004.

[173] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, July 2006.

[174] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *J. Mach. Learn. Res.*, 9:2491–2521, 2008.

[175] J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, 2009.

[176] Ben Calderhead and Mark Girolami. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus*, 1(6):821–835, 2011.

[177] Terry Speed. A Correlation for the 21st Century. *Science*, 334(6062):1502–1503, 2011.

[178] Rajan P Nair, Kristina Callis Duffin, Cynthia Helms, Jun Ding, Philip E Stuart, David Goldgar, Johann E Gudjonsson, Yun Li, Trilokraj Tejasvi, Bing-Jian Feng, Andreas Ruether, Stefan Schreiber, Michael Weichenthal, Dafna Gladman, Proton Rahman, Steven J Schrodi, Sampath Prahalad, Stephen L Guthery, Judith Fischer, Wilson Liao, Pui-Yan Kwok, Alan Menter, G Mark Lathrop, Carol A Wise, Ann B Begovich, John J Voorhees, James T Elder, Gerald G Krueger, Anne M Bowcock, and Gonalo R Abecasis. Genome-wide scan reveals association of psoriasis with il-23 and nf-κb pathways. *Nature Genetics*, 41(2):199–204, January 2009.

[179] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, 9(5):392–403, June 2008.

[180] Charles X. Ling, Jin Huang, and Harry Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pages 329–341, Acapulco, Mexico, 2003.

[181] Yvan Saeys, Thomas Abeel, et al. Robust feature selection using ensemble feature selection techniques. In *ECML PKDD*, pages 313–325, 2008.

[182] Yvan Saeys et al. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[183] Adam Woznica, Phong Nguyen, and Alexandros Kalousis. Model mining for robust feature selection. In *KDD*, pages 913–921, Beijing, 2012.

[184] Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast SVM training on very large data sets. *J. Mach. Learn. Res.*, 6:363–392, 2005.

[185] Bernhard Schölkopf et al. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001.

[186] Feiping Nie et al. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Trans. Neural Netw.*, 22(11):1796–1808, 2011.

[187] Itamar Arel, Derek C. Rose, and Thomas P. Karnowski. Deep Machine Learning - A New Frontier in Artificial Intelligence Research. *IEEE Comput. Intell. Mag.*, 5(4):13–18, November 2010.

[188] Xianshun Chen, Yew-Soon Ong, Meng-Hiot Lim, and Kay Chen Tan. A Multi-Facet Survey on Memetic Computation. *IEEE Trans. Evol. Comput.*, 15(5):591–607, October 2011.

[189] Christopher King and David A. Pendlebury. Research Fronts 2013. Technical report, April 2013.

[190] Ryan Meuth, Meng-Hiot Lim, Yew-Soon Ong, and Donald C. Wunsch. A Proposition on Memes and Meta-Memes in Computing for Higher-Order Learning. *Memetic Comp.*, 1(2):85–100, 2009.

[191] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 193–200, Corvallis, Oregon, United States, 2007.

[192] Sinno Jialin Pan, James T. Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd Association for the Advancement of Artificial Intelligence (AAAI 2008)*, volume 2, pages 677–682, Chicago, Illinois, United States, 2008. AAAI Press.

[193] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008)*, pages 353–360, Vancouver, B.C., Canada, December 2008. Curran Associates, Inc.

[194] James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.