<table>
<thead>
<tr>
<th>Title</th>
<th>The Emerging &quot;Big Dimensionality&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Zhai, Yiteng; Ong, Yew-Soon; Tsang, Ivor W.</td>
</tr>
<tr>
<td>Date</td>
<td>2014-08</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/39675">http://hdl.handle.net/10220/39675</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<a href="http://dx.doi.org/10.1109/MCI.2014.2326099">http://dx.doi.org/10.1109/MCI.2014.2326099</a>].</td>
</tr>
</tbody>
</table>
Abstract—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. Today, there is a growing consensus among the computational intelligence communities that data volume presents an immediate challenge pertaining to the scalability issue. However, when addressing volume in Big Data analytics, researchers in the data analytics community have largely taken a one-sided study of volume, which is the “Big Instance Size” factor of the data. The flip side of volume which is the dimensionality factor of Big Data, on the other hand, has received much lesser attention. This article 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 (variables) brings about new challenges to computational intelligence. We begin with an analysis on the origins of Big Dimensionality. The evolution of feature dimensionality in the last two decades is then studied using popular data repositories considered in the data analytics and computational intelligence research communities. Subsequently, the state-of-the-art feature selection schemes reported in the field of computational intelligence are reviewed to reveal the inadequacies of existing approaches in keeping pace with the emerging phenomenon of Big Dimensionality. Last but not least, the “curse and blessing of Big Dimensionality” are delineated and deliberated.

Index Terms—Big dimensionality, big data, high dimensionality, feature selection, survey

I. 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 we continue to create quintillion bytes daily [1]–[3], 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 5Vs in Table I [4]–[6].

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 our mind when we refer 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 [4]–[6], which refers to the “Big instance size” factor of the data; the corresponding factor of “Big Dimensionality”, on the other hand, has received much lesser attention in the context of Big Data analytics [8]. To date, some studies on high dimension small sample size problems have been reported, such as random projection, Naive Bayes, and others [9]–[12]. Theoretical efforts on Big Data with millions of dimensions have, however, remained relatively under-explored. In contrast to previous studies, in this article, we attempt 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, we begin with a peek through various stages of feature life cycle, namely feature description, feature selection and feature evaluation, as depicted in Figure 1. For each stage, we present a brief overview that focuses on the flip side of volume in Big Data – by putting spotlight on the emerging phenomenon of Big Dimensionality and the challenges ahead.

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 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 we 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 [13]. This myriad of feature types and their mixture are becoming a norm in many of today’s real world applications [14] and together with the rapid advancements in computing and information technologies, they are major contributors of feature explosion and continue to fuel Big Dimensionality.

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.

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 [17]. 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 [18]. Further, to verify the authenticity of the identified features for the application domains of interest, the 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. However, traditional feature verification and visualization approaches are likely to become obsolete in the face of Big Dimensionality.

Contributions: From our 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 paper is specifically set out to fulfill such a role. We begin 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.

1Both the IEEE International Conference on Data Mining (ICDM) 2013 and IEEE World Congress on Computational Intelligence (WCCI) 2014, organize workshops/special sessions that focus on the the issues of high dimensionality [15], [16], 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”.

2In the past, various new feature descriptors are inspired by feedbacks obtained from the verified models and visualized results [19], [20].
II. THE ORIGIN OF BIG DIMENSIONALITY

In this section, we focus on the key factors that accounts for the origin of Big Dimensionality. We begin with a study on the origin of features, 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, we showcase the domain of image and video learning, since it is a popular domain of computer science as motivated by the rising popularity of the Internet and mobile devices. Subsequently, we present 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.

A. 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 [21], 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 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. Figure 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 we 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.

B. 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 [22].

Today, the escalation of users that enjoy spending their leisure time watching videos, browsing photos and sharing

---

\(^3\)http://www.usa.canon.com/cusa/consumer/standard_display/Lens_Advantage_1S


\(^5\)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.

\(^6\)https://www.facebook.com

\(^7\)https://www.flickr.com
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, we discussed some of the core feature descriptors of online video content that contribute to the explosion of data dimensionality.

Figure 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 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 \textit{Spatial envelope}, for instance, was introduced as a holistic descriptor that captures the core objects in the pictures \cite{24}, see Figure 3B-I. \textit{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 \cite{25},

\(^8\)http://www.youtube.com/
see Figure 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 scene [26]. 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 3B-III. Subsequent extensions of SIFT include the Speeded up robust features (SURF) [27] and the Gradient speeded up robust features (SURF) [27] and the Gradient location and orientation histogram (GLOH) [28] descriptors, which were designed for gains in speed and prediction accuracy, respectively. Other popular image descriptors include the Quick shift [29] and the Difference of gaussians cornerness measure [30], etc., whose feature representations are depicted in Figure 3B-IV and Figure 3B-V, respectively.

In addition to the image features, Sub-figures 3C, 3D and 3E depict the other typical forms of descriptors considered in online video learning [31]–[34], 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 before moving on to in vitro study [35], [36]. The rapid advancements in biotechnologies and biodevices, nevertheless, has given researchers the option of using Single-Nucleotide Polymorphism (SNP) (see Figure 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 we now have 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.

C. The Evolution of Feature Dimensions in Data-Centric Research

In this subsection, we present an analysis on the developments of feature dimensions by focusing on three popular computational intelligence and data analytics repositories, namely, UC Irvine Machine Learning Repository (UCI) [37], UCI KDD Archive (UCI KDD) [38] and LIBSVM Database (LIBSVM) [39], that have transpired in the last two decades. Further, we consider a collection of high dimensional datasets from the three repositories, whose detailed characteristics including the year of creation, dimension size, data name and application domain, are tabulated in Table II.

**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

<table>
<thead>
<tr>
<th>Application Domain</th>
<th>Data Name</th>
<th>Dimension</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustics</td>
<td>ISOLET</td>
<td>617</td>
<td>1994</td>
</tr>
<tr>
<td>game</td>
<td>Chess</td>
<td>36</td>
<td>1989</td>
</tr>
<tr>
<td></td>
<td>Connect-4</td>
<td>42</td>
<td>1995</td>
</tr>
<tr>
<td>Image</td>
<td>Letter</td>
<td>16</td>
<td>1991</td>
</tr>
<tr>
<td></td>
<td>Corel</td>
<td>89</td>
<td>1999</td>
</tr>
<tr>
<td>Life Science</td>
<td>Soybean</td>
<td>35</td>
<td>1987</td>
</tr>
<tr>
<td></td>
<td>Molecular</td>
<td>58</td>
<td>1990</td>
</tr>
<tr>
<td></td>
<td>Mammals</td>
<td>72</td>
<td>1992</td>
</tr>
<tr>
<td></td>
<td>SPECTF</td>
<td>44</td>
<td>2001</td>
</tr>
<tr>
<td></td>
<td>PS3</td>
<td>5,409</td>
<td>2010</td>
</tr>
<tr>
<td>Multi-view</td>
<td>Internet AD</td>
<td>1,558</td>
<td>1998</td>
</tr>
<tr>
<td>Physics</td>
<td>Spectrometer</td>
<td>102</td>
<td>1993</td>
</tr>
<tr>
<td></td>
<td>Hg O Treat. Plant</td>
<td>38</td>
<td>2000</td>
</tr>
<tr>
<td>Sociology</td>
<td>Insurance</td>
<td>86</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>Bag of words</td>
<td>100,000</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>URL</td>
<td>3,231,961</td>
<td>2009</td>
</tr>
<tr>
<td>Time-series</td>
<td>PEEMS-SF</td>
<td>138,672</td>
<td>2011</td>
</tr>
<tr>
<td></td>
<td>Gas Sensor</td>
<td>1,950,000</td>
<td>2013</td>
</tr>
<tr>
<td>Video</td>
<td>YouTube MOV</td>
<td>1,000,000</td>
<td>2013</td>
</tr>
</tbody>
</table>

**LIBSVM Database**

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

End of sentence

end, we begin with a review on the state-of-the-art feature selection methods in this section. Subsequently, we proceed to reveal the core emerging challenges of feature selection when facing Big Dimensionality. Lastly, the “blessing of Big dimensionality” is also discussed.

A. Review of State-of-the-art Methods

Over the decades, a vast variety of feature selection methods have been proposed for handling different learning scenarios. Based on their learning paradigms, one can divide them into three categories – Filter methods, Wrapper methods and Embedded methods. Further, as feature correlation has been widely established as an important component of both feature and feature group selection, a subsection is also devoted to the topic.

1) Filter Methods: One simple way to select features is based on the ranking of features subject to some criteria. Gain Ratio [40] and Chi-Square [41] are among some of the earliest simple yet popular filter methods ever to be introduced. They rank features differently based on the criterion of information entropy or chi-square statistics. Such approaches however pay little attention on the feature interactions (i.e., only the predictive information of the individual feature is considered, which is often referred to as feature-level score). Correspondingly, on applications such as microarray data analysis [42], [43] where genes possess much intrinsic linkages, such approaches often suffer from the presence of suboptimality (local optima).

To address such problems, instead of the simple feature-level score, direct measurements on the contributions of feature subsets have been considered. One such effect of subset-level score methods is the general graph-based feature selection framework under a trace ratio criterion [44]. As graph depicts the relationships among data in a natural and effective way, weighted undirected graphs can be readily used to incorporate the within-class and between-class information. Another notable correlation based filter is minimum Redundancy Maximum Relevance (mRMR) [45], which selects the correlated

III. 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 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, we begin with a review on the state-of-the-art feature selection methods in this section. Subsequently, we proceed to reveal the core emerging challenges of feature selection when facing Big Dimensionality. Lastly, the “blessing of Big dimensionality” is also discussed.

A. Review of State-of-the-art Methods

Over the decades, a vast variety of feature selection methods have been proposed for handling different learning scenarios. Based on their learning paradigms, one can divide them into three categories – Filter methods, Wrapper methods and Embedded methods. Further, as feature correlation has been widely established as an important component of both feature and feature group selection, a subsection is also devoted to the topic.

1) Filter Methods: One simple way to select features is based on the ranking of features subject to some criteria. Gain Ratio [40] and Chi-Square [41] are among some of the earliest simple yet popular filter methods ever to be introduced. They rank features differently based on the criterion of information entropy or chi-square statistics. Such approaches however pay little attention on the feature interactions (i.e., only the predictive information of the individual feature is considered, which is often referred to as feature-level score). Correspondingly, on applications such as microarray data analysis [42], [43] where genes possess much intrinsic linkages, such approaches often suffer from the presence of suboptimality (local optima).

To address such problems, instead of the simple feature-level score, direct measurements on the contributions of feature subsets have been considered. One such effect of subset-level score methods is the general graph-based feature selection framework under a trace ratio criterion [44]. As graph depicts the relationships among data in a natural and effective way, weighted undirected graphs can be readily used to incorporate the within-class and between-class information. Another notable correlation based filter is minimum Redundancy Maximum Relevance (mRMR) [45], which selects the correlated

Fig. 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.

(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 archives9 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, we chart the dimensions of the datasets that has been used in the last two decades with respect to the year it was introduced. From Figure 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.

III. 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 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, we begin with a review on the state-of-the-art feature selection methods in this section. Subsequently, we proceed to reveal the core emerging challenges of feature selection when facing Big Dimensionality. Lastly, the “blessing of Big dimensionality” is also discussed.

A. Review of State-of-the-art Methods

Over the decades, a vast variety of feature selection methods have been proposed for handling different learning scenarios. Based on their learning paradigms, one can divide them into three categories – Filter methods, Wrapper methods and Embedded methods. Further, as feature correlation has been widely established as an important component of both feature and feature group selection, a subsection is also devoted to the topic.

1) Filter Methods: One simple way to select features is based on the ranking of features subject to some criteria. Gain Ratio [40] and Chi-Square [41] are among some of the earliest simple yet popular filter methods ever to be introduced. They rank features differently based on the criterion of information entropy or chi-square statistics. Such approaches however pay little attention on the feature interactions (i.e., only the predictive information of the individual feature is considered, which is often referred to as feature-level score). Correspondingly, on applications such as microarray data analysis [42], [43] where genes possess much intrinsic linkages, such approaches often suffer from the presence of suboptimality (local optima).

To address such problems, instead of the simple feature-level score, direct measurements on the contributions of feature subsets have been considered. One such effect of subset-level score methods is the general graph-based feature selection framework under a trace ratio criterion [44]. As graph depicts the relationships among data in a natural and effective way, weighted undirected graphs can be readily used to incorporate the within-class and between-class information. Another notable correlation based filter is minimum Redundancy Maximum Relevance (mRMR) [45], which selects the correlated

9http://archive.ics.uci.edu/ml/about.html
features that contribute most to the class labels such that they are mutually far apart from each other. This is achieved by maximizing the dependency between the joint distribution of the selected features and the output labels. In summary, methods that consider the interactions among features have received very interesting and notable results on small dimensional data, but generally succumb under applications involving millions of features due to computations on feature interactions. One such example is the aforementioned YouTube MVG dataset found in the UCI repository. In general, many empirical studies have shown that filter methods do not scale well above tens of thousands of features.

2) Wrapper and Embedded Methods: When dealing with Big Dimensionality, scalability is a primary concern. Here we review some of the notable wrapper and embedded methods that have been shown to handle large feature size, i.e., tens of thousands features, elegantly. Among them, Support Vector Machine (SVM) has been established to operate well on high dimensional classification problems [46]. Naturally, it is expected for numerous schemes using SVM for feature selection [47], [48] to showcase respectable prediction performance. For example, as a widely-studied wrapper approach in many different domains, the SVM based on recursive feature elimination (SVM-RFE) [36] operates by searching through all features individually, and eliminates the less important features iteratively. To solve high dimensional data efficiently, $\ell_1$-norm regularizer can be embedded into the SVM [49]–[51]. More recently, the feature generating machine [52] and group discovering machine [53] have been introduced with notable results on datasets comprising millions of dimensions. These frameworks operate with the cutting plane algorithm and multiple kernel learning to make coping with millions of dimensions feasible.

3) Feature Correlation: From our survey, feature correlation plays an essential role in filter, wrapper and embedded methods, as demonstrated in the context of biomarker networks [54] and natural language processing [13]. In particular, feature correlation plays a crucial part of feature group selection in computational intelligence. The graph-guided fused lasso is among the early approaches of feature group selection approaches and operates by identifying feature groups based on the graph-structure define over the features [55]. Octagonal shrinkage and clustering algorithm for regression, on the other hand, incorporate the $\ell_\infty$-penalty so as to reduce similar feature pairs [56]. Further, Zhong et al. introduced an efficient projection step to accelerate the process of feature grouping [57]. With a non-convex optimization formulation [58], some amount of bias alleviation can be also achieved.

B. 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 II, 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, we begin with a discussion on some inadequacies of current computational intelligence methodologies, as they were not designed to cope with Big Dimensionality. Consequently, we highlight 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.

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, we cite the biomarker feature selection problem in life science as the illustrating example. The search for a compact subset of relevant biomarkers [59] 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 [35], [36], [60]. 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, we conducted an analysis on the dimensionality of datasets that have been used in the studies of computational intelligence, which is summarized in Figure 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 and Learning Systems (TNN10) 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 5), the complexity of the feature selection tasks being addressed began to overwhelm the algorithms proposed (see Figure 6) to date. Particularly, the dimensionality of the algorithms under-studied (as summarized in Figure 6) is significantly lagging behind those being produced (see Figure 5). From Figure 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 [13], [111], [114]. In summary, it is becoming clear that the explosion of dimensionality is pushing the capability limits of current computational intelligence algorithms.

2) Handling Trillion Correlations: With millions of features in hand, existing computational intelligence approaches

---

10Before 2012, this transactions was named as “IEEE Transactions on Neural Networks”.
that require the calculations of pairwise correlations in their algorithmic designs will have to cope with computations in the range of trillions\textsuperscript{11}. For example, a dataset with millions of features ($10^9$) 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.

3) 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

\textsuperscript{11}The number pairwise correlation computations is a squared of the dimensionality.
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.

4) Visualization That Matters: 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 [118] as “the science of analytical reasoning facilitated by interactive visual interfaces”. A presentation of the diabetes dataset given in Figure 7(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 7(b) remains legible, although the correlation coefficients and feature labels can no longer be included with ease. In the case of the psoriasis12 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. Thus, as the feature size continues to grow and evolve towards the 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.

C. Blessing of Big Dimensionality

Besides the challenges (curse), in this section we review the potential benefits that are attributed by the presence of Big Dimensionality, which is less widely noted than the former [119].

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 8. 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 8 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.

IV. CONCLUDING REMARKS

In this article, 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. 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.

Through this article, 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. Hence, 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 [120], Memetic Computation [121] or otherwise, for improved scalability, applicability and reusability. These atomized units of memes, metamemes, or memeplex can then be expressed in hierarchical nested relationships or conceptual entities for higher-order learning [123], thus forming societies of the mind for more effective problem solving.

ACKNOWLEDGMENT

This work is partially supported under the A*STAR-TSRP funding, the Singapore Institute of Manufacturing Technology - Nanyang Technological University (SIMTech-NTU) Joint Laboratory and Collaborative Research Programme on Complex Systems, and the Centre for Computational Intelligence (C²I) at NTU.

REFERENCES


[20] Y. Sun and B. Bhanu, “Image Retrieval with Feature Selection and


