Invited Paper

Feature Selection in Bioinformatics

Lipo Wang[a]

[a]School of Electrical and Electronic Engineering, Nanyang Technological University, Block S1, 50 Nanyang Avenue, Singapore 639798

ABSTRACT

In bioinformatics, there are often a large number of input features. For example, there are millions of single nucleotide polymorphisms (SNPs) that are genetic variations which determine the difference between any two unrelated individuals. In microarrays, thousands of genes can be profiled in each test. It is important to find out which input features (e.g., SNPs or genes) are useful in classification of a certain group of people or diagnosis of a given disease. In this paper, we investigate some powerful feature selection techniques and apply them to problems in bioinformatics. We are able to identify a very small number of input features sufficient for tasks at hand and we demonstrate this with some real-world data.

Keywords: Bioinformatics, feature selection, SNP, microarray

1. INTRODUCTION

We are facing a dramatic increase in data available in many domains due to advances in technologies. For example, trillions of web links represent an explosion in web data, notably because of the popularity in social networks. In bioinformatics, various large projects, such as the human genone project, together with new techniques, such as the microarray, have created anormous amount of data. These data often come with high dimensionality. For example, web documents have thousands of words and microarray data can involve tens of thousands of genes. Such high data dimensionality can significantly increase computational burden, even to the extend that it render some data mining approaches impossible. For example, it would be very difficult to train a neural network or support vector machines with tens of thousands input nodes. Furthermore, many of these input features are irrelevant to a given task and can act like noise to decrease performance. Feature selection, which finds a small set of input features for a problem at hand, thus has paramount importance.[1–6]

This paper discusses two specific examples of feature selection in bioinformatics, i.e., selecting relevant single nucleotide polymorphisms (SNPs) to determine certain groups of people and selecting significant genes in microarray data for diagnosis of a given disease. We demonstrate our approaches with some real-world data.

2. FEATURE SELECTION FOR THE HAPMAP GENOTYPE DATA

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations and determine human diversities, e.g., different physical traits, different predispositions to diseases, and different responses to medicine. It is thus important (1) to find out which set of SNPs differentiate individuals into different population groups and (2) to be able to accurately classify individuals to different population groups using these relevant SNPs.

Selection algorithms for informative SNPs (namely tag SNPs) in association studies[15,22,37] were based on correlation among SNPs, e.g., the linkage disequilibrium (LD)[8,23] measure γ^2.[10,24] The purpose was to find out how those selected tag SNPs predict or represent other SNPs. In population studies, SNPs are selected to classify different populations, therefore, tag SNP selection methods will be different from those in association studies. Rosenberg et al. (2003)[26] proposed to use the informativeness for assignment (I_n) to measure the ability of each genetic loci or marker (feature) to infer individuals' ancestry, which is proved to be similar to the F-statistics measure.[26] In 2005, Rosenberg et al.[25] proposed four algorithms, i.e., exhaustive, univariate, greedy and maximum, to select marker panels with performance approaching the maximum. The four algorithms were realized through a given performance function, i.e., the optimal rate of correct assignment (ORCA),[25] which measures the probability of correctly assigning an individual to the population from which the genotype of the individual has originated with the greatest possibility.

Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X, edited by Harold Szu, Liyi Dai, Proc. of SPIE Vol. 8401, 840113 · © 2012 SPIE · CCC code: 0277-786X/12/$18
doi: 10.1117/12.921417

In this section, we select a set of tag SNPs which should lead to the best classification accuracy. Different from previous algorithms,[15,22,25,26,37] we first use a feature importance ranking measure, i.e., a modified t-test, to rank each SNP (the input feature) according to its discriminative capability. Secondly, according to the ranking list, we greedily choose different SNP subsets with different numbers of SNPs, say 5, 10, 50, 100 and so on, and test them on a classifier, e.g., the support vector machine (SVM).[30,31] The proper feature subset (tag SNP subset) is the one with the highest classification accuracy and minimum size.

The most common t-test, i.e., the student t-test,[11] may be used to assesses whether the means of two classes are statistically different from each other by calculating a ratio between the difference of two class means and variability of the two classes. The t-test has been used to rank features (genes) for microarray data[18,27] and for mass spectrometry data.[20,36] Those applications of t-test are only limited to 2-class problems. For multi-class problems,[28] calculated a t-statistic value (Equations (1)) for each gene of each class by evaluating the difference between the mean of one class and the mean of all the classes, where the difference is standardized by the within-class standard deviation.

$$t_{ic} = \frac{\overline{x}_{ic} - \overline{x}_i}{M_c \cdot (S_i + S_0)} \tag{1}$$

$$S_i^2 = \frac{1}{N-C} \sum_{c=1}^{C} \sum_{j \in c} (x_{ij} - \overline{x}_{ic})^2 \tag{2}$$

$$M_c = \sqrt{1/n_c + 1/N} \tag{3}$$

Here t_{ic} is the t-statistics value for gene (feature) i of class c. $\overline{x}_{ic}$ is the mean of feature i in class c and $\overline{x}_i$ is the mean of feature i for all classes. x_{ij} is feature i of sample j. N is the number of all the samples in the C classes and n_c is the number of samples in class c. S_i is the within-class standard deviation and S_0 is set to be the median value of S_i for all the features.[28] used the t-statistics to shrink class means toward the mean of all classes to constitute a nearest shrunken centroid classifier and did not mention how to use the t-statistic value to rank genes with regard to all the classes.[32] extended the t-score for feature i to be the greatest t-score for all classes for feature i:

$$t_i = max \left\{ \frac{|\overline{x}_{ic} - \overline{x}_i|}{M_c S_i}, c = 1, 2, ...C \right\} \tag{4}$$

The SNP data available at www.hapmap.org has nominal components, for example, AA, AT and TG. The existing t-statistic do not handle nominal data. Therefore, We generalize the t-score of each feature as follows:

1. Suppose the feature set is $F = \{f_1, ..., f_i, ..., f_g\}$, and feature i has m_i different nominal values represented as $f_i = \{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m_i)}\}$
2. Transform each nominal feature value into a vector with the dimension m_i, i.e., $x_i^{(1)} \Rightarrow \vec{X}_i^{(1)}$=(0, ..., 0, 1), $x_i^{(2)} \Rightarrow \vec{X}_i^{(2)}$=(0, ..., 1, 0), ..., $x_i^{(m_i)} \Rightarrow \vec{X}_i^{(m_i)}$=(1, ..., 0, 0).
3. Replace all the numerical features in Equations (1) and (2) with those vectors (see Equations (5) and (6)).

$$t_i = max \left\{ \frac{\left| \overline{\vec{X}}_{ic} - \overline{\vec{X}}_i \right|}{M_c S_i}, c = 1, 2, ...C \right\} \tag{5}$$

$$S_i^2 = \frac{1}{N-C} \sum_{c=1}^{C} \sum_{j \in c} (\vec{X}_{ij} - \overline{\vec{X}}_{ic})(\vec{X}_{ij} - \overline{\vec{X}}_{ic})^T \tag{6}$$

We rank the features as follows: the greater the t-scores, the more relevant the features.

Table 1. Classification accuracy (standard deviation) for each of the original 4 classes.

Number of SNPs	CEU	CHB	JPT	YRI
5	66.68%(20.2%)	39.51%(36.77%)	21.06%(28.7%)	88.20%(7.78%)
10	72.14%(17.42%)	33.86%(22.64%)	39.34%(27.85%)	93.17%(5.28%)
50	82.67%(12.64%)	43.74%(23.14%)	55.95%(19.71%)	99.85%(0.82%)
100	88.57% (8.5%)	54.59%(21.72%)	49.10%(20.0%)	100% (0)
200	95.98%(5.54%)	56.88%(20.25%)	54.95%(19.78%)	100% (0)
300	98.76%(3.21%)	59.76%(24.17%)	55.18%(24.59%)	100% (0)
400	99.67%(1.22%)	58.42%(23.61%)	55.84%(18.78%)	100% (0)
500	99.86%(0.75%)	59.09%(21.52%)	50.06%(21.59%)	100% (0)
1000	99.40%(1.87%)	58.34%(23.37%)	51.40%(26.76%)	100% (0)

The support vector machine (SVM)[30,31] has been extensively applied in bioinformatics. Compared with many traditional machine learning approaches, the SVM shows significantly better or at least matched performance.[17] We choose the SVM in our experiment to test different feature subsets and find the best discriminative feature subsets, i.e., the one with the best classification accuracy and the minimum size.

The genotype data in the HapMap database (www.hapmap.org) includes four populations, i.e., CEU, YRI, JPT and HCB. Here CEU represents Utah residents with ancestry from northern and western Europe. YRI represents Yoruba individuals from Ibadan and Nigeria. Each of the two populations has 90 reference individuals (samples) which are comprised of 30 father-mother-offspring trios. JPT represents Japanese individuals from Tokyo, and HCB means Han Chinese individuals from Beijing. Each of these two populations has 45 samples and the individuals in each of the populations are unrelated. For CEU and YRI samples, we remove the children samples so that all the samples are unrelated. Thus the total number of samples used in our experiment is 210. We will carry out classification with the 4 populations groups.

Most data samples are strings of bi-allelic SNPs, i.e., with each SNP feature containing only two alleles. Few of SNP features have 3 or more alleles at each position, which are called multi-allelic, and will be omitted in this paper as.[15,37] Because some populations do not have any information provided at some SNP positions, we remove those SNP positions (features) and obtain nearly 4 million SNPs for our experiment. For each bi-allelic SNP feature, there are at most three feature types (values). For example, if two alleles that constitute a feature are the same, e.g., A and A, there will be only AA for this feature, which is known as homozygous. Otherwise, two different alleles, e.g., A and T, will constitute three feature types: AA, AT and TT, which is called hyterozygous. Therefore, for the three nominal values of each feature, we will use 3 vectors with dimension 3 to represent them, e.g., (0, 0, 1), (0, 1, 0), and (1, 0, 0). For the feature with the homozygous type, one numeric value is enough to represent it.

With the results of feature ranking, we use the greedy selection[19] to form different feature subsets with different sizes for classification. Since the number of features is so large that we can not handle all the data simultaneously due to memory constraint in the computer, we deal with one chromosome at a time. We first rank features in each chromosome, separately. Then we combine the 22 ranking lists for the 22 chromosomes together and rank again to obtain the total ranking list, from which we selected 5, 10, 50, 100, 200, 300, 400, 500 and 1000 top features to form 9 different feature subsets. For each feature subset, the training and testing are run 30 times by the SVM. Each time we randomly choose 140 samples as the training set and 70 as the testing set.

We choose the radial basis function (RBF) kernel for the SVM. The kernel parameter and the penal parameter are decided by cross-validation and grid search method.[16] Classification results are shown in Table 1.

From the table we can see that when the number of features increases, the average classification accuracy gradually increases. With the top 400 features obtained from the t-test ranking measure, the accuracy reaches the maximum of 81% on average. Hence among the original nearly 4 million SNPs, only a minority of them,

e.g., 400 or so, are very important for differentiating the populations and most of the 4 million SNPs may be redundant or irrelevant.

3. GENE SELECTION IN MICROARRAY DATA

Based on gene expression obtained from microarrays,[38] cancers can be classified into appropriate sub-types using various machine learning and statistical methods, such as artificial neural network,[39] evolutionary algorithm,[40] nearest shrunken centroids.[41] In particular, Tibshirani et al. successfully classified lymphoma data set[42] with only 48 genes by using a statistical method called nearest shrunken centroids with an accuracy of 100%.[43] In this section, we show how to find 2 genes to achieve an accuracy of 100%.

In the lymphoma data set[42] (http://llmpp.nih.gov/lymphoma), there are 42 samples derived from diffuse large B-cell lymphoma (DLBCL), 9 samples from follicular lymphoma (FL), 11 samples from chronic lymphocytic leukaemia (CLL). The entire data set includes the expression data of 4026 genes. In this data set, a small part of data are missing. A k-nearest neighbor algorithm was applied to fill those missing values.[44]

We randomly divided the 62 samples into 2 parts, 31 samples for training, 31 samples for testing.

We first attempted to classify the data set using only 1 gene by training and testing our fuzzy neural network (FNN).[45–47] The best 5-fold CV accuracy was 90.32% and the best testing accuracy was 80.65%. We then tried all possible combinations of 2 genes within the 100 genes with the highest t-scores. Among all 4950 such possible 2-gene combinations, the CV accuracy for the training data reached 100% in 174 combinations. The corresponding testing accuracies for these combinations varied from 80.6% (6 errors in 31 testing samples) to 100%.

We also tried the method in the SRBCT data set[39] (http://research.nhgri.nih.gov/microarray/Supplement) which contains the expression data of 2,308 genes. There are 63 training samples and 25 testing samples. There are 4 classes, i.e., Ewing family of tumors (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and Burkitt lymphomas (BL). We found a 3-gene set that can achieve 100% 5-fold cross-validation classification accuracy.

For a large andcomplex data set with 14 cancer types, we divided the whole problem into a group of binary classification problems and applied t 2-step approach to each of these binary classification problems. The GCM data set includes 14 types of cancers[48] (http://wwwgenome. wi.mit.edu/mpr/GCM.html). Through this divide-and-conquer approach, we obtained accura comparable to previously reported results but with only 28 genes rather than 16,063 genes.

4. CONCLUSIONS

In this paper, we demonstrated the effectiveness of feature selection in bioinformatics through 2 particular problems, i.e., population classification with SNPs and cancer classification with microarray data. We proposed to find out which SNPs are significant in determining the population groups and to classify different populations using these relevant SNPs as the input features. We proposed a modified t-test ranking measure based on those discussed in[28,32] and applied it on the problem of classifying populations from the Hapmap genotype data.

It is very important to realize population classification with few SNPs. The significance of this work can be viewed from two aspects. From a computational point of view, it would be much cheaper to handle several hundred SNPs rather than the original 10 million common SNPs directly. Some of the SNPs may be irrelevant and therefore act as "noise" to tasks of classification and clustering. From a biological point of view, reducing the number of irrelevant SNPs can facilitate geneticists to focus on fewer SNPs, so as to reduce genotyping cost and increase efficiency of association studies and population studies.

At the same time, we should notice that for this application we only did a coarse feature selection (greedy selection of feature subsets after feature ranking). We did not detect feature correlations in order to remove those redundant features. In our future work, we will deal with those redundant features by calculating correlations among features or by clustering. Furthermore, we will also try to form novel feature combinations, in which selected features need not be the most highly ranked and the size of the feature subset can be further reduced.[21,32]

Similarly, we showed how to reduce the number of genes needed to achieve the highest possible classification accuracy from tens of thousands to much smaller numbers, for example, 2 in the Lymphoma data set, 3 in the

SRBCT data set, and 28 in the GCM data set. Our future work will attempt to apply our powerful feature selection approaches to other important bioinformatics problems.

REFERENCES

[1] X.J. Fu and L.P. Wang, "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance", *IEEE Trans. System, Man, Cybern, Part B - Cybernetics,* vol.33, no.3, pp. 399-409, 2003.

[2] Bing Liu, Chunru Wan, and L.P. Wang, "An efficient semi-unsupervised gene selection method via spectral biclustering", *IEEE Transactions on Nano-Bioscience*, vol.5, no.2, pp.110-114, June, 2006.

[3] Nina Zhou and L.P. Wang, "Effective selection of informative SNPs and classification on the HapMap genotype data", *BMC Bioinformatics*, 8:484, 2007.

[4] L.P. Wang, Nina Zhou, and Feng Chu, "A general wrapper approach to selection of class-dependent features," *IEEE Transactions on Neural Networks*, vol.19, no.7, pp.1267-1278, 2008.

[5] F. Chu and L.P. Wang, "Applications of support vector machines to cancer classification with microarray data," *International Journal of Neural Systems*, vol.15, pp. 475-484, 2005.

[6] X.J. Fu and L.P. Wang, "A GA-based novel RBF classifier with class-dependent features," *Cec'02: Proceedings of the 2002 Congress on Evolutionary Computation*, vols.1 and 2, pp. 1890-1894, 2002.

[7] Avi-Itzhak, *et al.* 2003. Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity, *In Proc. of Pac. Symp. Biocomput.*, 8: 466-477.

[8] Celedon, J. C. 2004. Candidate genes, SNPs, Haplotypes and linkage disequilibrium. Powerpoint presentation.

[9] Daly, M. J., *et al.* 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.*, 29: 229-232.

[10] Devlin, B. and Risch, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping, *Genomics*, 29: 311-322.

[11] Devore, J. and Peck, R. 1997. Statistics:the exploration and analysis of data, 3rd ed. Pacific Grove, CA: Duxbury Press.

[12] Guyon, I. and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 3: 1157-1182.

[13] Halgamuge, S. and Wang, L. P. (eds.) 2005. Classification and Clustering for Knowledge Discovery, Springer, Berlin.

[14] Halgamuge, S. and Wang, L. P. (eds.) 2005. Computational Intelligence for Modeling and Predictions, Springer, Berlin.

[15] Halperin, E., *et al.* 2005. Tag SNP selection in genotype data for maximizig SNP prediction accuracy, *Bioinformatics*, 199: 195-203.

[16] Hsu, *et al.* 2003. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei.

[17] Hua, S. and Sun, Z. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, *Journal of Molecular Biology*, 308: 397-407.

[18] Jaeger, J., *et al.* 2003. Improved Gene Selection For Classification Of Microarrays, *Pac. Symp. Biocomput.*, pp. 53-64.

[19] Kwak, N. and Choi, C. H. 2002. Input Feature Selection by Mutual Information Based on Parzen Window, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24: 1667-1671.

[20] Levner, I. 2005. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics*, 6: 68, doi: 10.1186/1471-2105-6-68

[21] Liu, B., *et al.* 2006. An efficient semi-unsupervised gene selection method via spectral biclustering, *IEEE Trans. on Nano-Bioscience*, 5: 110-114.

[22] Liu, T. F., *et al.* 2005. Effective Algorithms for Tag SNP Selection, *Journal of Bioinformatics and Computational Biology*, 3: 1089-1106.

[23] Phuong, T. M., *et al.* 2005. Choosing SNPs using Feature Selection, *Proc IEEE Comput Syst Bioinform Conf. (CSB'05)*, pp. 301-309.

[24] Pritchard, J. K. and Przeworski, M. 2001. Linkage disequilibrium in humans: models and data, *Am. J. Hum. Genet.*, 69: 1-14.

[25] Rosenberg, N. A. 2005. Algorithms for selecting informative marker panels for population assignment, *Journal of computational biology*, 12: 1183-1201

[26] Rosenberg, N. A. *et al.* 2003. Informativeness of Genetic Markers for Inference of Ancestry, *Am. J. Hum. Genet.*, 73: 1402-1422.

[27] Su, Y., *et al.* 2003. RankGene: Identifcation of Diagnostic Genes Based on Expression Data. *Bioinformatics*, 19: 1578-1579.

[28] Tibshirani, R., *et al.* 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99: 6567–6572.

[29] Trochim, W. M. 2004. The Research Methods Knowledge Base, 2nd Edition, Atomic Dog Publishing.

[30] Vapnik, V. 1998. Statistical learning theory, Wiley, NewYork.

[31] Wang, L. P. 2005. Support Vector Machines: Theory and Applications, Springer.

[32] Wang, L. P., Chu, F. and Xie, W. 2007. Accurate cancer classification using expressions of very few genes, *IEEE Transactions on Bioinformatics and Computational Biology*, 4: 40-53.

[33] Wang, L. P. and Fu, X. J. 2005. Data Mining with Computational Intelligence, Berlin: Springer-Verlag.

[34] Welch, B. L. 1947. The generalizaition of student's problem when several different population are involved, *Biomethika*, 34: 28-35.

[35] Wright, S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating, *Evolution*, 19: 395-420.

[36] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K. and Zhao, H. 2003. Comparison of statistical methods for classifcation of ovarian cancer using mass spectrometry data. *BioInformatics*, 19: 1636-1643.

[37] Zhen, L. and Altman, R. B. 2004. Finding Haplotype Tagging SNPs by Use of Principle Components Analysis, *Am. J. Hum. Genet.*, 75: 850-861.

[38] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray, Science, vol. 270, pp. 467-470, 1995.

[39] J.M. Khan et al., Classification and Diagnostic Prediction o Cancers Using Gene Expression Profiling and Artificial Neural Networks, Nature Medicine, vol. 7, pp. 673-679, 2001.

[40] J. Deutsch, Evolutionary Algorithms for Finding Optimal Gen Sets in Microarray Prediction, Bioinformatics, vol. 19, pp. 45-52, 2003.

[41] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, Diagnosis o Multiple Cancer Types by Shrunken Centroids of Gene Expression, Proc. Natl Academy of Sciences USA, vol. 99, pp. 6567-6572, 2002.

[42] A.A. Alizadeh et al., Distinct Types of Diffuse Large b-Cell Lymphoma Identified by Gene Expression Profiling, Nature, vol. 403, pp. 503-511, 2000.

[43] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, Class Predicition by Nearest Shrunken Centroids with Applications to DNA Microarrays, Statistical Science, vol. 18, pp. 104-117, 2003.

[44] O. Troyanskaya et al., Missing Value Estimation Methods for DNA Microarrays, Bioinformatics, vol. 17, pp. 520-525, 2001.

[45] L.P. Wang and Yakov Frayman, "A dynamically generated fuzzy neural network and its application to torsional vibration control of tandem cold rolling mill spindles", *Engineering Applications of Artificial Intelligence*, vol.15, no.6, pp. 541-550, 2002.

[46] Y. Frayman and L.P. Wang,"Data mining using dynamically constructed recurrent fuzzy neural networks," *Research and Development in Knowledge Discovery and Data Mining*, vol. 1394, pp. 122-131, 1998.

[47] Y. Frayman and L.P. Wang, "A Dynamically-constructed fuzzy neural controller for direct model reference adaptive control of multi-input-multi-output nonlinear processes," *Soft Computing*, vol.6, pp.244-253, 2002.

[48] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub, Diagnosis Using Tumor Gene Expression Signature, Proc. Natl Academy of Sciences USA, vol. 98, pp. 15 149-15 154, 2000.