Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 6, Number 2 (2013) pp. 81-86

http://www.ripublication.com/jcib.htm

Gene Expression with Phenotype Classification and Patient Survivel Prediction using Datamining Technologies

[1]T. Shanmugavadivu and [2]T. Ravichandran

[1]Research Scholar ,Karpagam University, Coimbatore
[2]Principal, Hindusthan Institute of Technology, Coimbatore

Abstract

The past few decades witness an explosive growth in biological information generated by the scientific community. This is caused by major advances in the field of molecular biology, coupled with advances in genomic technologies. In turn, the huge amount of genomic data generated not only leads to a demand on the computer science community to help store, organize and index the data, but also leads to a demand for specialized tools to view and analyze the data. The main role of bioinformatics was to create and maintain databases to store biological information, such as nucleotide and amino acid sequences. With more and more data generated, nowadays, the most pressing task of bioinformatics has moved to analyze and interpret various types of data, including nucleotide and amino acid sequences, protein domains, protein structures and so on. To meet the new requirements arising from the new tasks, researchers in the field of bioinformatics are working on the development of new algorithms (mathematical formulas, statistical methods and etc) and software tools which are designed for assessing relationships among large data sets stored.

Keywords : Bagging decision tree, optimal data set, supervised learning.

1. INTRODUCTION

At the end of the 1980's a new discipline, named data mining, emerged. The introduction of new technologies such as computers, satellites, new mass storage media and many others have lead to an exponential growth of collected data. Traditional data analysis techniques often fail to process large amounts of -often noisy- data efficiently, in an exploratory fashion. The scope of data mining is the

knowledge extraction from large data amounts with the help of computers. It is an interdisciplinary area of research, that has its roots in databases, machine learning, and statistics and has contributions from many other areas such as information retrieval, pattern recognition, visualization, parallel and distributed computing. There are many applications of data mining in real world. Customer relationship management, fraud detection, market and industry characterization, stock management, medicine, pharmacology, and biology are some examples.

Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large scale experiments and research programs. This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology. The main aims of bioinformatics are:

- The organization of data in such a way that allows researchers to access existing information and to submit new entries as they are produced.
- The development of tools that help in the analysis of data.
- The use of these tools to analyze the individual systems in detail, in order to gain new biological insights.

2. MOTIVATION

The aim of data mining is to automatically or semi-automatically discover hidden knowledge, unexpected patterns and new rules from data. There are a variety of technologies involved in the process of data mining, such as statistical analysis, modeling techniques and database technology. During the last ten years, data mining is undergoing very fast development both on techniques and applications. Its typical applications include market segmentation, customer profiling, fraud detection, (electricity) loading forecasting, and credit risk analysis and so on. In the current post-genome age, understanding floods of data in molecular biology brings great opportunities and big challenges to data mining researchers. Successful stories from this new application will greatly benefit both computer science and biology communities.

3. LITERATURE REVIEW

This kind of learning is a process from general to specific and is supervised because the class membership of training instances is clearly known. In contrast to supervise learning is unsupervised learning, where there are no pre-defined classes for training instances. The main goal of unsupervised learning is to decide which instances should be grouped together, in other words, to form the classes. Sometimes, these two kinds of learning's are used sequentially — supervised learning making use of class information derived from unsupervised learning. This two-step strategy has achieved some success in gene Table 1.1: An example of gene expression data. There are two samples, each of which is described by 5 genes. The class label in the last column indicates the phenotype of the sample.

Table 1.1 Gene Table

Gene1	Gene2	Gene3	Gene4	Gene5	Class
298	654	1284	800	163	ALL
2947	1811	198	679	225	AML

expression data analysis field, where unsupervised clustering methods were first used to discover classes (for example, subtypes of leukemia) so that supervised learning algorithms could be employed to establish classification models and assign a phenotype to a newly coming instance. Feature weighting algorithms assign weights to features individually and rank them based on their relevance to the target concept. There are a number of different definitions on feature relevance in machine learning literature. A feature is good and thus will be selected if its weight of relevance is greater than a threshold value.

A well known algorithm that relies on relevance evaluation is Relief. The key idea of Relief is to estimate the relevance of features according to how well their values distinguish between the instances of the same and different classes that are near each other. Relief randomly samples a number (m) of instances from the training set and updates the relevance estimation of each feature based on the difference between the selected instance and the two nearest instances of the same and opposite classes. Time complexity of Relief for a data set with M instances and N features is O(mMN). With m being a constant, the time complexity becomes O(MN), which makes it very scalable to data sets with both a huge number of instances and a very high dimensionality.

However, Relief does not help with removing redundant features. As long as features are deemed relevant to the class concept, they will all be selected even though many of them are highly correlated to each other. Many other algorithms in this group have similar problems as Relief does. They can only capture the relevance of features to the target concept, but cannot discover redundancy among features. However, empirical evidence from feature selection literature shows that, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms and thus should be eliminated as well. Therefore, in the context of feature selection for high dimensional data where there may exist many redundant features, pure relevance-based feature weighting algorithms do not meet the need of feature selection very well.

Subset search algorithms search through candidate feature subsets guided by a certain evaluation measure which captures the goodness of each subset. An optimal (or near optimal) subset is selected when the search stops. Some existing evaluation measures that have been shown effective in removing both irrelevant and redundant features include the consistency measure and the correlation measure. Consistency measure attempts to find a minimum number of features that separate classes as consistently as the full set of features can. An inconsistency is defined as two instances having the same feature values but different class labels. In Dash et al.,

different search strategies, namely, exhaustive, heuristic, and random search, are combined with this evaluation measure to form different algorithms. The time complexity is exponential in terms of data dimensionality for exhaustive search and quadratic for heuristic search. The complexity can be linear to the number of iterations in a random search, but experiments show that in order to find best feature subset, the number of iterations required is mostly at least quadratic to the number of features. In Hall, a correlation measure is applied to evaluate the good -ness of feature subsets based on the hypothesis that a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other. The underlying algorithm, named CFS, also exploits heuristic search. Therefore, with quadratic or higher time complexity in terms of dimensionality, existing subset search algorithms do not have strong scalability to deal with high dimensional data.

To overcome the problems of algorithms in both groups and meet the demand for feature selection for high dimensional data, we develop a novel algorithm which can effectively identify both irrelevant and redundant features with less time complexity than subset search algorithms.

4. DECISION TREE

Decision tree induction is among the most popular classification methods. As mentioned above, decision tree has an important advantage over other machine learning algorithms such as k-NN and SVM, in a qualitative dimension: rules produced by decision tree induction are easy to interpret and understand, and hence, can help greatly in appreciating the underlying mechanisms that separate samples in different classes. In general, decision trees try to find an optimal partitioning of the space of possible observations, mainly by the means of subsequent recursive splits. Most of the algorithms implement this induction process in a top-down manner: (1) determining the root feature that most discriminatory with regard to the entire training data; (2) using the root feature to split the data into non-overlapping subsets; (3) selecting a significant feature of each of these subsets to recursively partition them until reaching one of stopping criteria.

4.1.1 Bagging of decision trees

The technique of bagging was coined by Breiman, who investigated the properties of bagging theoretically and empirically for both classification and numeric prediction. Bagging of trees combines several tree predictors trained on bootstrap samples of the training data and gives prediction by taking majority vote. In bagging, given a training set S with n samples, a new training set S_0 is obtained by drawing n samples uniformly with replacement from S. When there is a limited amount of training samples, bagging attempts to neutralize the instability of single decision tree classifier by randomly deleting some samples and replicating others.

Algorithm for bagging

Generation of trees:

- Let n be the number of samples in the training data S.

- For each of k iterations:
- Obtain a new training set S_0 by drawing n samples with replacement from S.
- Apply the decision tree algorithm to S_0 .Store the resulting tree.

Classification:

- Given a new sample.
- For each of the k trees:
- Predict class of sample according to the tree.
- Return class that has been predicted most often.

5. CONCLUSION

In the aspect of classification algorithms, no single algorithm is absolutely superior to all others, though SVM achieves fairly good results in most of tests. Compared with SVM, decision tree methods can provide simple, comprehensive rules and are not very sensitive to feature selections. Among the decision tree methods, the newly implemented CS4 achieves good prediction performance and provides many interesting rules.

Feature generation is important for some kinds of biological data. The researcher illustrates this point by properly constructing new feature space for functional sites recognition in DNA sequences. Some of the signal patterns identified from the generated feature space is highly consistent with related literature or biological knowledge. The rest might be useful for biologists to conduct further analysis.

REFERENCES:

[1] P. Agarwal and V. Bafna. The ribosome scanning model for translation initiation: implicationsfor gene prediction and full-length cDNA detection. *Proceedings of 6th Internationa lConference on Intelligent Systems for Molecular Biology*, pages 2–7, June1998.

[2] Y. Aissouni, C. Perez, B. Calmels, and P.D. Benech. The cleavage/polyadenylation activity triggered by a U-rich motif sequence is differently required depending on the poly(A)site location at either the first or last 3’-terminal exon of the 2’-5’ oligo(A) synthetase gene. *Journal of Biological Chemistry*, 277:35808–35814, 2002.

[3] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C.Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick,H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J.Jr. Hudson, L. Lu,D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger,J.O. Armitage, R. Warnke, and L.M. Staudt. Distinct types of diffuse large B-cell lymphomaidentified by gene expression profiling. *Nature*, 403:503–511, 2000.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of

tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of National Academy of Sciences of the United States of American*, 96:6745–6750, 1999.

[5] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D.Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, 2002.

[6] Ghosh, D. and Chinnaiyan, A.M. (2002). Mixture Modeling of Gene Expression Data from Microarray Experiments. Bioinformatics, 18, 275-286.

[7] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, Molecular Classification of Cancer: Class Discovery and Class Predictionby Gene Expression Monitoring. Science 286(5439), 531-537.