ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

SURVEY ON DATA MINING ALGORITHMS TO PREDICT LEUKEMIA TYPES

[1] B. RAJESWARI,
[1] M.Phil Research Scholar,
[1] PG & Research Department of Computer Science,

[2] ARUCHAMY RAJINI,
[2] Associate Professor,
[2] PG & Research Department of Computer Application,

[1, 2] Hindusthan College of Arts & Science, Coimbatore.

Abstract:-

Data mining is defined as shifting through very large amounts of data for useful in formation. Some of the most important and popular data mining techniques are association rules, classification, clustering, prediction and sequential patterns. Data mining techniques are used for variety of applications. In health care industry, data mining plays an important role for predicting diseases. Recent advances in microarray technology offer the ability to measureexpression levels of thousands of genes simultaneously. Analysis of such data helpsus identifying different clinical outcomes that are caused by expression of a fewpredictive genes.. The feature extraction and classification are carried outwith combination of the high accuracy of ensemble based algorithms, and comprehensibility of a single decision tree. These allow deriving exact rules by describinggene expression differences among significantly expressed genes in leukemia. It isevident from our results that it is possible to achieve better accuracy in classifyingleukemia without sacrificing the level of comprehensibility.

Keywords: - [Prediction Algorithms, Data mining in cancer cell, leukemia, mining classifiers]

1 INTRODUCTION

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and data-base systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been de-fined as “the nontrivial extraction of previously unknown, implicit and potentially useful information from data. It is “the science of extracting useful information from large databases”. It is one of the tasks in the process of knowledge discovery from the database. Data Mining is used to discover knowledge Out of data and presenting it in a form that is easily understood to humans. Clinical diagnosis for disease prediction is one of the most important emergingapplications of microarray gene expression study. In the last decade, a newtechnology, DNA microarrays, has allowed screening of biological samples fora huge number of genes by measuring expression patterns. This technologyenables the monitoring of the expression levels of a large portion of a genomeon a single slide or "chip", thus allowing the study of interactions among thousands of genes simultaneously. Usually microarray datasets are usedfor identification of differentially expressed genes, which from data miningpoint of view represents a feature selection problem. The objectives of thisresearch are to select important features from leukemia predictive genes andto derive a set of rules that classify differentially expressed genes. The studyfollows the comprehensibility of a single decision tree. Although, there aremany research that have demonstrated a higher level of accuracy in classifyingcancer cells,, the comprehensibility issue of decision treesto gain best accuracy in the domain of microarray data analysis has beenignored .In this study, we attempt to combine the high accuracy of ensembles andthe interpretability of the single tree in order to derive exact rules that describe differences between significantly expressed genes that are responsiblefor leukemia. To achieve this, Combined Multiple Models (CMM) method has been applied, which was proposed originally by Domingosin.

In our study the method is adapted for multidimensional and real valued microarraydatasets to eliminate the co linearity and multivariate problems. All datasetsfrom our experiments are publicly available from the Kent Ridge Repository. These microarray samples are the examples of human tissue extracts that are related to a specific disease and have been used forcomprehensible interpretation in this study. The following sections explorethe datasets, methods of CMM adaptation and testing. It also presents theresults that are obtained by applying the adapted method on four publiclyavailable databases.

Finally the chapter presents a validation study by providing an interpretation of the results in the context of rule sets and then bycomparing the proposed adaptations with the combined and simple decisiontrees for leukemia grouping.

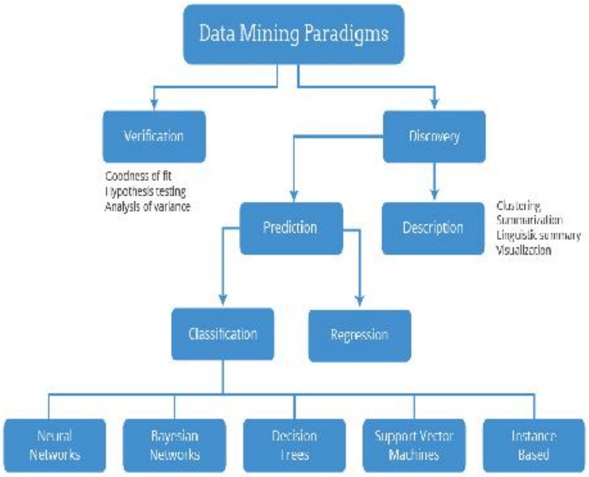

Figure 1: Data Mining Paradigms

2. Data cleaning

With the term of data cleaning we refer to the task of detecting and correctingor removing corrupt or inaccurate records from a dataset, before applyinga data mining algorithm. Microarray data cleaning includes the followingissues.Normalization. Normalization is needed to adjust the individual hybridizationintensities to balance them appropriately so that meaningful Biological comparisons can be made. It ensures that differences in intensities are due to differential expression and not some printing, hybridizationor scanning artifacts. Several normalization methods have been proposed in literature and some software package has been developedfor the analysis of microarray data. One of the most popular and generalpurpose software packages for microarray data is Bio conductor .Other software are distributed by the companiesthat produce the microarray technology, like Affymetrix and Agilent.

Missing value estimation. Missing values in microarray data arisedue to technical failures, low signal-to-noise ratio and measurement errors.For example, dust present on the chip, regularities in the spot productionand inhomogeneous hybridization all lead to missing values. It has beenestimated that typically 1% of the data are missing affecting up to 95% ofthe genes [81]. To limit the effects of missing values several works addressedthe problem of missing value estimation, and the most used approach is thek-nearest neighbor's algorithm.

Outlier detection. The problem of outliers defined as "anomalous datapoints" often arises in large datasets. The aim of outlier detection methodsis to detect and remove or substitute outliers. A broad survey of methods that have been found useful in the detection and treatment of outlierson microarray data analysis is done. Usually outliers are detectedby computing the mean and the standard deviation of values. Thevalues outside the range are considered outliers. Other techniques havebeen proposed by replacing the mean and the standard deviation values, forexample by using 3σ instead of σ. An alternative specifically used in themicroarray data analysis community is the Hampel identifier, which replaces the mean with the median and the standard deviation with the medianabsolute deviation.

3. USING CLASSIFICATION ALGORITHMS

The most used classification algorithms exploited in the microarray analysisbelong to four categories: decision tree, Bayesian classifiers, neural networksand support vector machines.Decision Tree. Decision tree derives from the simple divide-and-conqueralgorithm. In these tree structures, leaves represent classes and branchesrepresent conjunctions of features that lead to those classes. At each nodeof the tree, the attribute that most effectively splits samples into differentclasses is chosen. To predict the class label of an input, a path to a leaf fromthe root is found depending on the value of the predicate at each node that is visited An evolution of decision tree exploited for microarray data analysis is the random forest , which uses an ensemble of classificationtrees. It showed the good performance of random forest for noisy and multi-class microarray data.

3.1 Bayesian classifiers and Naive Bayesian.

From a Bayesian viewpoint,a classification problem can be written as the problem of finding theclass with maximum probability given a set of observed attribute values. Suchprobability is seen as the posterior probability of the class given the data, and is usually computed using the Bayes theorem. Estimating this probabilitydistribution from a training dataset is a difficult problem, because it mayrequire a very large dataset to significantly explore all the possible combinations.Conversely, Naive Bayesian is a simple probabilistic classifier basedon Bayesian theorem with the (naive) independence assumption. Based onthat rule, using the joint probabilities of sample observations and classes, thealgorithm attempts to estimate the conditional probabilities of classes givenan observation. Despite its simplicity, the Naive Bayes classifier is knownto be a robust method, which shows on average good performance in termsof classification accuracy, also when the independence assumption does not hold.

3.2 Artificial Neural Networks (ANN).

An artificial neural network is amathematical model based on biological neural networks. It consists of aninterconnected group of artificial neurons and processes information using aconnectionist approach to computation. Neurons are organized into layers.The input layer consists simply of the original data, while the output layer nodes represent the classes. Then, there may be several hidden layers. A key feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted in order to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained.

3.3 Support vector machines (SVM).

Support vector machines are a relatively new type of learning algorithm, originally introduced by . Intuitively, SVM aims at searching for the hyper plane that best separates the classes of data. SVMs have demonstrated the ability not only to correctly separate entities into appropriate classes, but also to identify instances whose established classification is not supported by data. Although SVMs are relatively insensitive to the distribution of training examples in each class, they may still get stuck when the class distribution is too skewed.

Sometimes, a combination of the presented methods may outperform the single technique. For example, a method which combines a neural network classifier. In order to consider the correlations among genes, they build a neural network where the weights are determined by a Bayesian method. A Bayesian approach combined with SVM to determine the separating hyper plane of an SVM, once its maximal margin is determined in the traditional way.

4. COMBINED MULTIPLE MODELS FOR GENE EXPRESSION ANALYSIS

Data mining is the process of autonomously extracting useful information or knowledge from large datasets. Many different models can be used in data mining process. However, it is required for many applications not only to involve model that produce accurate predictions, but also to incorporate comprehensible model. In many applications it is not enough to have accurate model, but we also want comprehensible model that can be easily interpreted to the people not familiar with data mining. For example, Tibshirani and Knight proposed a method called Bumping that tries to use bagging and produce a single classifier that best describes the decisions of bagged ensemble. It builds models from bootstrapped samples and keeps the one with the lowest error rate on the original data. Typically this is enough to get good results also on test set. We should also mention papers that suggest different techniques of extracting decision trees from neural networks or ensembles of neural networks that can all be seen as a “black-box” method .

4.1 Gene Selection

It has been shown that selecting a small subset of informative genes can lead to improved classification accuracy and greatly improves execution time of data mining tools . The most commonly used gene selection approaches are based on gene ranking. In these gene ranking approaches, each gene is evaluated individually and assigned a score representing its correlation with the class. Genes are then ranked by their scores and the top ranked ones are selected from the initial set of features (genes). To make our experiments less dependent of the filtering method, we use three different filtering methods. This way we get 12 different microarray datasets with a pre-defined number of most relevant gene expressions. All used filtering methods are part of WEKA toolkit that we were using in our experiments.

The following filtering methods were used:

4.2 Gain Ratio filter

This is the heuristic that was originally used by Quinlan in ID3 [27]. It is implemented in WEKA as a simple and fast feature selection method. The idea of using this feature selection technique for gene ranking was already presented by Ben-Dor et al. [28].

4.3 Relief-F filter

The basic idea of Relief-F algorithm [29] is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. A study comparing Relief-F to other similar methods in microarray classification domain was conducted by Wang and Makedon [30] where they conclude that the performance of Relief-F is comparable with other methods.

4.4 SVM filter

Ranking is done using Support Vector Machines (SVM) classifier. Similar approach using SVM classifier.

CONCLUSION

In this survey paper the problem of summarizing the different algorithms of data mining for the major life threatening diseases are used in the field of medical prediction are discussed. Acute leukemia which is of lymphoid origin is called Acute LymphocyticLeukemia (ALL) and a malignant disorder where myeloid blast cells accu-mulate in the marrow and bloodstream is called Acute MyelocyticLeukemia. A study conducted by Golub has revealed 50 predictivegenes that differentiate between ALL and AML. In a recentstudy conducted by Umpai and Aitken has demonstrated that the geneX95735 zyxin significantly determines AML whereas myosin light chain overexpression frequency is higher in ALL patients. Some other studies also havedemonstrated the similar result.

REFERENCES:

[1] J. Li and K. Ramamohanarao, A Tree-based Approach to the Discovery of Diagnostic Biomarkers for Ovarian Cancer, in Proceedings of the PAKDD 2004, pp. 682–691, Sydney, Australia, February 2004

[2] M. Dettling, Bag Boosting for tumor classification with gene expression data, Bioinformatics, vol. 20, no. 18, pp. 3583–3593, 2004

[3] D. P. Berrar, B. Sturgeon, I. Bradbury, C. S. Downes and W. Dubitzky, Mi- croarray Data Integration and Machine Learning Techniques For Lung Cancer Survival Prediction, in Proceedings of Critical Assessment of Microarray Data Analysis (CAMDA 2003), Durham, North Carolina, USA, pp. 43–54, November 2003

[4] P. Domingos, Knowledge discovery via multiple models, Intelligent Data Analysis, vol. 2 no. 1–4, pp. 187–202, 1998

[5] R. Tibshirani and K. Knight, Model search and inference by bootstrap bumping, Journal of Computational and Graphical Statistics, vol. 8, pp. 671–686, 1999

[6] O. Boz, Converting a Trained Neural Network To a Decision Tree DecText Decision Tree Etxractor, PhD thesis, Computer Science and Engineering,Lehigh University, 2000