International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 4, Issue 7 (November 2012), PP. 53-57

APRIORI-HYBRID ALGORITHM AS A TOOL FOR COLON CANCER MICROARRAY DATA CLASSIFICATION

Merry KP[1], Rabindra Kumar Singh[2], Swaroop.S.Kumar[3]
Hindustan Institute of Technology and Science[1,2], UniBiosys Biotech Research Labs [3]

The advent of microarray technology has enabled the researchers to rapidly measure the levels of thousands of genes expressed in a biological tissue sample in a single experiment. One important application of this microarray technology is to classify the tissue samples using their gene expression profiles, identifying several types of cancer. Association rules are one of the most important data mining concepts which can be defined as the relation and dependency between the itemsets by a given support and confidence in database. These itemsets consists of genes from gene expression data which are highly expressed or repressed. In this paper an attempt is made to classify benchmark colon cancer microarray dataset using Association rule mining algorithm, namely Apriori-Hybrid. Apriori-Hybrid, it is the combination of algorithm Apriori and Apriori-TID, which can classify the large itemsets and can improve the accuracy of classification of cancer and it can also shed light on the basic mechanism that enable each cancer type to survive and thrive, which inturn help in early detection of the type of cancer. We propose Apriori-Hybrid as an improvised algorithm for tumor classification.
Keywords-Microarray, Association rules, Apriori, Apriori-TID, Apriori-Hybrid

I. INTRODUCTION

DNA microarrays or alternative quantification techniques have enabled genome-wide expression analyses of various biological phenomena. The most important application of the microarray technique is to classify unknown samples according to their expression profile, e.g., to discriminate cancerous or noncancerous samples or to discriminate different types or subtypes of cancer. Cancer detection and classification for diagnostic and prognostic purposes is generally based on pathological analysis of tissue sections, resulting in subjective interpretations of data. The limited information gained from morphological analysis is often insufficient to aid in cancer diagnosis and may result in expensive but ineffective treatment of cancer. In order to accurately identify cancer subtypes, many recent studies have been carried out to identify genes that might cause cancer. Advances in microarray technology and improved methods for processing and deciphering biological data have augmented these studies.

The analysis is expected to overcome the conventional problems of histopathological cancer diagnosis such as variations in diagnosis by individual pathologists or difficulties in differentiating between malignant and benign tissues due to their morphological similarities. For constructing diagnosis systems using high-dimensional gene expression data, supervised learning theories are often applied, and several studies have been successful in recent years. Many classification methods originated from statistical learning theory have been adapted for molecular data classification or clustering.

Classification rule mining and association rule mining are two important data mining techniques. Classifications rule mining aims to discover a small set of rules in the database to form an accurate classifier (e.g., Quinlan 1992; Breiman et al. 1984). Association rule mining finds all rules in the databases that satisfy some minimum support and minimum confidence constraints (e.g., Agrawal and Srikant 1994). For association rule mining, the target of mining is not predetermined, while for classification rule mining there is one and only one predetermined target, i.e., the class. Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form LHS→RHS, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes**.** A formal statement of the association rule problem is [Agrawal1993] [Cheung1996c].

II. MATERIALS AND METHODS

2.1. Collection of Data

Benchmark colon tumor dataset was collected from Kent Ridge Repository, which contains 62 samples of colon cancer patients. Among them 40 tumor biopsies are from tumors (labeled as negative) and 22 normal (labeled as positive) biopsies are from healthy parts of the colon of the same patient.

2.2. Feature Selection

T-Test statistics was used as feature selection method for the selection of highly expressed genes. T-Test statistic is defined as follows to measure the weighted mean differences for feature X between the two classes of target Y.

$$\boldsymbol{t(X_i, Y)} = \frac{\overline{X_{i+}} - \overline{X_{i-}}}{\sqrt{s_{i+}/m_{+} + s_{i-}/m_{-}}}$$

Where m± is the number of the samples in class ±1 respectively, $\overline{X_i}$ ± and s± denote the sample mean and sample standard deviation of X for each class of Y.

2.3 Discretization

Defined a cut-off value where anything above this setting will be considered as up regulated('1') and anything below will be considered as down regulated('0').

2.4. Generation of association rules

Association rules were generated for the whole sample as well as for the positive and negative samples separately. The positive and negative samples association rules were used to train the association classifier.

The most highly expressed 250, 500, 1000 genes where selected from 2000 genes using T-Test statistics. On the basis of number of genes cut-off were taken separately, and genes were discretized into zeros and ones. After discretization whole support and individual gene support was calculated.

$$\textbf{Support} = \frac{count\ of\ \uparrow\ genes}{Array\ count}$$

The genes which were having support less than the whole support were pruned and the rest were taken into consideration. For the generation of rules confidence was set into 50% and confidence were calculated between the genes.

$$\textbf{Confidence} = \frac{Gene\ 1 \cup Gene\ 2}{Gene\ 1}$$

If the confidence is above 50% consider the combination and say there is a rule between these two genes and the below confidence combinations were pruned out. Association rules were generated until seven combinations using Apriori-Hybrid algorithm. Apriori-Hybrid is a combination of two algorithms namely,Apriori and Aprior-TID. Apriori in the initial passes and switches to Apriori-TID in later passes for the generation of rules.

2.5. Classification

Classifier was trained first using the positive and negative association rules to identify the positive and negative association rules from the whole sample. The set of rules of the whole sample that were selected after the pruning phase were given to the classifier after training the classifier. The association classifier gives the count of positive and negative rules and which indirectly helped in the prediction of the percentage of positive and negative sample in the whole sample.

III. ALGORITHMS

3.1 Association Rule Mining

Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form LHS→RHS, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes. A formal statement of the association rule problem is [Agrawal1993] [Cheung1996c]:

Definition 1: Let I = {I1, I2, ..., Im} be a set of m distinct attributes, also called literals. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$ .An association rule is an implication of the form $X \Rightarrow Y$, where X, $Y \subset I$, are sets of items called itemsets, and $X \cap Y = \varphi$. Here, X is called antecedent, and Y consequent.

Two important measures for association rules, support (s) and confidence (α), can be defined as follows.

Definition 2: The support (s) of an association rule is the ratio (in percent) of the records that contain $X \cup Y$ to the total number of records in the database.

Definition 3: For a given number of records, confidence (α) is the ratio (in percent) of the number of records that contain $X \cup Y$ to the number of records that contain X

.

3.2 Apriori Algorithm

The Apriori algorithm developed by [Agrawal1994] is a great achievement in the history of mining association rules [Cheung1996c]. It is by far the most well-known association rule algorithm. This technique uses the property that any subset of a large itemset must be a large itemset. Also, it is assumed that items within an itemset are kept in lexicographic order. The fundamental differences of this algorithm from the AIS and SETM algorithms are the way of generating candidate itemsets and the selection of candidate itemsets for counting. As mentioned earlier, in both the AIS and SETM algorithms, the common itemsets between large itemsets of the previous pass and items of a transaction are obtained. These common itemsets are extended with other individual items in the transaction to generate candidate itemsets. However, those individual items may not be large. As we know that a superset of one large itemset and a small itemset will result in a small itemset, these techniques generate too many candidate itemsets which turn out to be small.

The Apriori algorithm addresses this important issue. The Apriori generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database. By only considering large itemsets of the previous pass, the number of candidate large itemsets is significantly reduced.

In the first pass, the itemsets with only one item are counted. The discovered large itemsets of the first pass are used to generate the candidate sets of the second pass using the apriori_gen() function. Once the candidate itemsets are found, their supports are counted to discover the large itemsets of size two by scanning the database. In the third pass, the large itemsets of the second pass are considered as the candidate sets to discover large itemsets of this pass. This iterative process

terminates when no new large itemsets are found. Each pass *i* of the algorithm scan the database once and determine large itemsets of size *i*. Li denotes large itemsets of size *i*, while Ci is candidates of size *i*.

Algorithm Apriori [Agrawal1994]

Input: I, D,
Output: L
Algorithm:
//Apriori Algorithm proposed by Agrawal R., Srikant, R. [Agrawal1994]
//procedure LargeItemsets
1) C 1: = I; //Candidate 1-itemsets
2) Generate L1 by traversing database and counting each occurrence of an attribute in a transaction;
3) for (k = 2; Lk-1≠ φ; k++) do begin
//Candidate Itemset generation
//New k-candidate itemsets are generated
from (k-1)-large itemsets
4) Ck = apriori-gen(Lk-1);
//Counting support of Ck
5) Count (Ck, D)
6) Lk = {c∈ Ck | c.count ≥ minsup}
7) end
8) L:= Uk Lk

3.3. Apriori-TID

The apriori-gen function to determine the candidate itemsets before the pass begins. The interesting feature is that the database D is not used for counting support after the first pass. Rather, the set |Ck| is used for this purpose. Each member of the set |Ck| is of the form < TID; {Xk}>, where each Xk is a potentially large k-itemset present in the transaction with identifier TID. For k = 1, C1 corresponds to the database D, although conceptually each item i is replaced by the itemset {i}. For k > 1, |Ck| is generated by the algorithm. The member of Ck corresponding to transaction t is <t.TID, {cÎ Ck|c contained in t}>. If a transaction does not contain any candidate k-itemset, then |Ck| will not have an entry for this transaction. Thus, the number of entries in |Ck| may be smaller than the number of transactions in the database, especially for large values of k. In addition, for large values of k, each entry may be smaller than the corresponding transaction because very few candidates may be contained in the transaction. However, for small values for k, each entry may be larger than the corresponding transaction because an entry in Ck includes all candidate k-itemsets contained in the transaction.

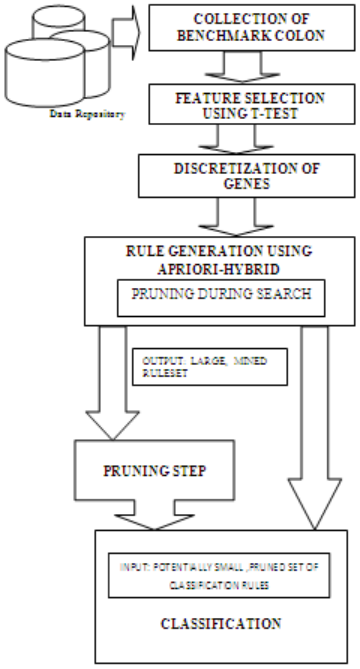
Algorithm AprioriTid

1) L1 = {large 1-itemsets};
2) C1 = database D;
3) for (k = 2; Lk-1 ≠∅ ; k++) do begin
4) Ck = apriori-gen(Lk-1);
5) |Ck| = ∅;
6) forall entries t ∈ |Ck-1| do begin
7) // determine candidate itemsets in Ck
contained in the transaction with identifier
t.TID
Ct = {c 2∈Ck | (c - c[k]) ∈t.set-of-itemsets ^ (c
– c [k-1]) ∈ t.set-of-itemsets};
8) forall candidates c ∈ Ct do
9) c.count++;
10) if (Ct≠∅) then |Ck| += < t.TID,Ct >;
11) end
12) Lk = {c ∈ Ck | c.count >= minsup}
13) end
14) Answer = ∪k Lk;

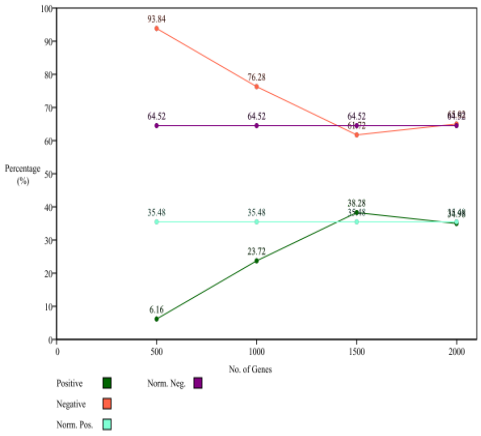
3.4. Apriori-Hybrid

This algorithm is based on the idea that it is not necessary to use the same algorithm in all passes over data. As mentioned in [Agrawal1994], Apriori has better performance in earlier passes, and Apiori-TID outperforms Apriori in later passes. Based on the experimental observations, the Apriori-Hybrid technique was developed which uses Apriori in the initial passes and switches to Apriori-TID when it expects that the set Ck at the end of the pass will fit in memory. Therefore, an estimation of Ck at the end of each pass is necessary. Also, there is a cost involvement of switching from Apriori to Apriori-TID. The performance of this technique was also evaluated by conducting experiments for large datasets. It was observed that Apriori-Hybrid performs better than Apriori except in the case when the switching occurs at the very end of the passes. Apriori-Hybrid is being used for the cancer classification as mentioned above.

IV. ARCHITECTURE



EXPERIMENTAL RESULTS



V. CONCLUSION

Apriori-Hybrid, it is the combination of algorithm Apriori and Apriori-TID, which can classify the large itemsets and can improve the accuracy of classification of cancer and it can also shed light on the basic mechanism that enable each cancer type to survive and thrive, which inturn help in early detection of the type of cancer. We propose Apriori-Hybrid as an improvised algorithm for tumor classification.

REFERENCES

[1] S.Ghorai,A.Mukherjee and P.K.Dutta, ”Cancer Classification from Gene Expression Data by NPPC Ensemble”, IEEE/ACM Transactions On Computational Biology and Bioinformatics,vol.8,No.3,May/June 2011.

[2] Yukyee Leung and Yeungsam Hung,”A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification”,IEEE/ACM Transactions On Computational Biology and Bioinformatics, vol. 7, no. 1, January/March 2010.

[3] Topon Kumar Paul and Hitoshi Iba, "Prediction of cancer class with majority voting genetic programming classifier using gene expression data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 6, No. 2, April-June 2009.

[4] I. Lonnstedt and T. Britton, "Hierarchical Bayes Models for CDNA Microarray Gene Expression," Biostatistics, vol. 6, no. 2, pp. 279-291, 2005.

[5] W. Chu, Z. Ghahramani, F. Falciani, and D. Wild, "Biomarker Discovery in Microarray Gene Expression Data with Gaussian Processes," Bioinformatics, vol. 21, no. 16, pp. 3385-3393, 2005.

[6] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, and B.K. Mallick, "Gene Selection: A Bayesian Variable Selection Approach," Bioinformatics, vol. 19, no. 1, pp. 90-97, 2003.

[7] X. Zhou, X. Wang, and E.R. Dougherty, "Gene Prediction Using Multinomial Probit Regression with Bayesian Gene Selection," EURASIP J. Applied Signal Processing, vol. 1, pp. 115-124, 2004.

[8] X. Zhou, K. Liu, and S.T.C. Wong, "Cancer Classification and Prediction Using Logistic Regression with Bayesian Gene Selection", J. Biomedical Informatics, vol. 37, pp. 249-259, 2004.

[9] Xin Zhao and Leo Wang-Kit Cheung," Multiclass Kernel-Imbedded Gaussian Processes for Microarray Data Analysis",IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 4, July/August 2011.

[10] Chun-Hou Zheng, Lei Zhang, To-Yee Ng, Simon C.K. Shiu, and De-Shuang Huang, " Metasample-Based Sparse Representation for Tumor Classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 5, September/October 2011.

[11] Naoto Yukinawa, Shigeyuki Oba, Kikuya Kato, and Shin Ishii , "Optimal Aggregation of Binary Classifiers for Multiclass Cancer Diagnosis using Gene Expression Profiles", Transactions on Computational Biology and Bioinformatics , vol. 6, No. 2, April-June 2009.

[12] Saras Saraswathi, Suresh Sundaram, Narasimhan Sundararajan, Michael Zimmermann, and Marit Nilsen-Hamilton,"ICGA-PSO-ELM Approach for Accurate Multiclass Cancer Classification Resulting in Reduced Gene Sets in Which Genes Encoding Secreted Proteins Are Highly Represented", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 2, March/April 2011.

[14] Tara McIntosh and Sanjay Chawla, "High Confidence Rule Mining for Microarray Analysis", Transactions on Computational Biology and Bioinformatics, vol. 4, No. 4, October-December 2007.

[15] Margaret H. Dunham, Yongqiao Xiao Le Gruenwald and Zahid Hossain,"A Survey Of Association Rules". International Journel of Computer TheoryAnd Engineering, vol.4, No.2 June 2003

[16] M. K. Ghoseand K. Gauthaman, "Association Rule Mining in Genomics", International Journel of Computer TheoryAnd Engineering, vol.2, No.2 April 2010.