Biological Data Prediction Using Two Mode Grouping Bayesian Principal

M. Sangeetha, P. Bhuvaneswari, A. Sujitha, P. Nandhini, C. Gurulakshmi

Abstract*— *The development of DNA chip technology makes it possible that high-throughput gene expression profiles could be observed simultaneously in particular living organism. The obtained data are usually shown in the form of matrix with genes in rows and experimental conditions in columns. However, these matrices often contain missing values caused by various factors, such as hybridization failures, insufficient resolution, or deposition of dust or scratches on the slide. The subsequent analyses of gene expression data (e.g. clustering, inferring regulatory model, or finding functional gene) always require the complete matrices. Repeating the experiments to obtain a complete gene expression matrix is usually costly and unpractical. Omitting the gene expression profile vector with missing values may lose useful information. Substituting the missing values with zeros or row averages lead the change of variance among variables. So an efficient imputation method for the missing value is needed.

***Index Terms*— DNA Chip, Hybridization, Clustering, Genes**

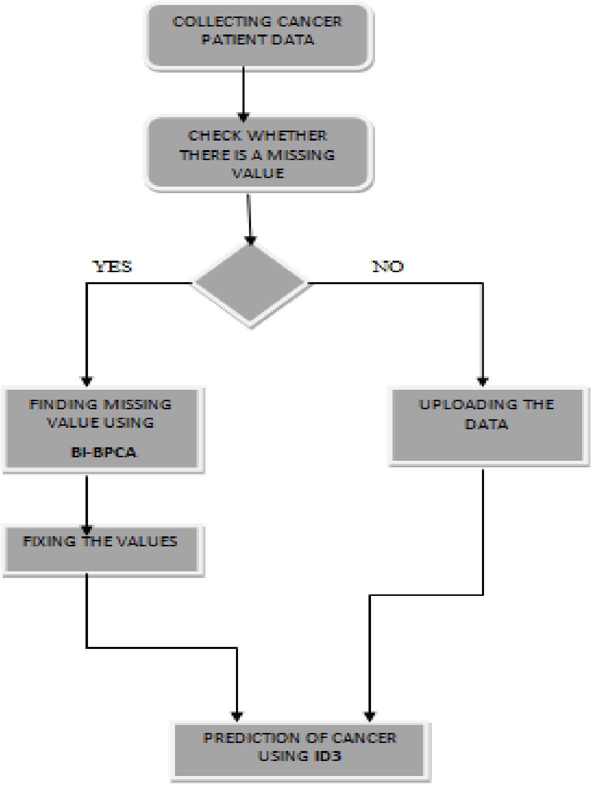
I. INTRODUCTION

Migration of patients record from one place to another may lead to some missing of data. Missing values are often encountered in gene expression Genes may show the similarity not in the whole conditions but in a certain subspace. Since the current methods consider the global information resulting in high computational cost and the distance criterion may bring less accurate results, Although the complexity of the biclustering problem may depend on the exact problem formulation, and, specifically, on the merit function used to evaluate the quality of a given bicluster, almost all interesting variants of this problem are NP-complete. Gene expression matrices have been extensively analyzed in two dimensions: the gene dimension and the condition dimension. This correspond to the Analysis of expression patterns of genes by comparing rows in the matrix. Analysis of expression patterns of samples by comparing columns in the matrix. Misssing of data which is found by bayesian method is not much satisfactory. so we are going for Bi-cluster BPCA method . Biclusters are coherent clusters consisting of correlated genes(rows) under correlated experimental conditions (columns). In this paper, these correlated experimental conditions are the columns that are correlated with the missing entry. The concept of biclustering was introduced early, but it did not become popular until 2000 when Cheng and Church applied it in the gene expression matrices introduced a geometrical biclustering method where biclusters embedded in a matrix can be regarded as points distributed on special linear structures in high-dimensional space, and the Hough transform is applied to find these linear patterns in the high-dimensional space so that biclusters can be recognized.

Manuscript published on 28 February 2015.
***** Correspondence Author (s)

Ms. M. Sangeetha, AP/IT Department, Sri Krishna College of Technology, Coimbatore, India.
Ms. P. Bhuvaneswari, AP/IT Department, Sri Krishna College of Technology, Coimbatore, India.
Ms. A. Sujitha, AP/IT Department, Sri Krishna College of Technology, Coimbatore, India.
Ms. P. Nandhini, AP/IT Department, Sri Krishna College of Technology, Coimbatore, India.
Ms. C. Gurulakshmi, AP/IT Department, Sri Krishna College of Technology, Coimbatore, India.

II. OVERALL DIAGRAM


COLLECTING CANCER PATIENT DATA
CHECK WHETHER THERE IS A MISSING VALUE
YES
NO
FINDING MISSING VALUE USING BI-BPCA
UPLOADING THE DATA
FIXING THE VALUES
PREDICTION OF CANCER USING ID3

III. EXISTING SYSTEM

Bayesian principal component analysis is a well known microarray missing value estimation method. A bi cluster based BPCA can be used for finding the missing values. Data generated from microarray experiments often suffer from missing value. As most downstream analyses need full matrices as input, these missing values have to be estimated. In Bicluster , the most correlated genes and experimental conditions with the missing entry are identified, and BPCA is conducted on these biclusters to estimate the missing values.

Decision tree learning, used in statistics , data mining and machine learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Decisiong tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorisation and generalisation of a given set of data. Decision graphs have been further extended to allow for previously unstated new attributes to be learnt dynamically and used at different places within the graph.

IV. PROPOSED SYSTEM

Proposed Method: Bi-Cluster Bayesian Principal

Cancer is one of the deadliest diseases among many people across the world. Our project aims at helping the medical practitioners to diagnose the patients at the early stage which can reduce the number of deaths. The decision tree is an important classification method in data mining classification. The proposed work is that we have modified the id3 algorithm using decision tree classification method and included the pre processing steps for the cancer data set to improve the accuracy of the classifier. The data set has missing values in it. In the pre processing steps of the data set, we have resolved it. Also the data set has data conflicts in it. And we have proposed an approach to resolve it. The concept of conditional entropy measure is used in the id3 algorithm and modified it. Then after pre processing the data set, it is supplied to the modified algorithm which constructs the decision tree and thus it proves to increase the accuracy of the classifier.

Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. We examine the decision tree learning algorithm ID3 and implement this algorithm using C# programming. We first implement basic ID3 in which we dealt with the target function that has discrete output values. We also extend the domain of ID3 to real valued output, such as numeric data and discrete outcome rather than simply Boolean value. The Java applet provided at last section offers a simulation of decision-tree learning algorithm in various situations. Some shortcomings are discussed in this project as well.

V. ADVANTAGES OF PROPOSED SYSTEM

Amongst other data mining methods, decision trees have various advantages

Simple to understand and interpret

People are able to understand decision tree models after a brief explanation.

Requires little data preparation

Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed.

Able to handle both numerical and categorical data

Other techniques are usually specialized in analysing datasets that have only one type of variable. (For example, relation rules can be used only with nominal variables while neural networks can be used only with numerical variables.)

Uses a white box model

If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. (An example of a black box model is an artificial neural network since the explanation for the results is difficult to understand.)

Possible to validate a model using statistical tests

That makes it possible to account for the reliability of the model.

Robust

Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

Performs well with large datasets

Large amounts of data can be analysed using standard computing resources in reasonable time.

VI. OVERVIEW OF WORK

1. FINDING MISSING VALUE

2. FIXING MISSED VALUE

3. FINDING ACCURACY

4. PREDICTING CANCER DISEASE

FINDING MISSING VALUE

In prediction of cancer disease , each and every data of patient record is important. so when migration take place from one hospital to another hospital some data may lead to miss. that can be found and can be fixed using bi bpca algorithm.

FIXING MISSED VALUE

The value is missed due to the migration of database so the missed value is fixed by using matrix format in the form of gene and attributes . it can be found by creating row and column matrix and finding the accurate value.

FINDING ACCURACY

The missed value is fixed by using the matrix and it is compared to the null value. for finding the accuracy we need to assign the null value so that it is compared and the accurate value is found.

PREDICTING CANCER DISEASE

Finally the missed value will be fixed using decision tree algorithm in the form of matrix format and by using that missed value and graph will be generated. by this cancer occurrence can be predicted.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

VII. METHODOLOGY

Attribute Selection

How does ID3 decide which attribute is the best. A statistical property, called information gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. In order to define gain, we first borrow an idea from information theory called entropy. Entropy measures the amount of information in an attribute.

Given a collection S of c outcomes

Entropy(S) = S -p(I) log2 p(I)

where p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.

Note that S is not an attribute but the entire sample set.

Example 1

If S is a collection of 14 examples with 9 YES and 5 NO examples then

Entropy(S) = - (9/14) Log2 (9/14) - (5/14) Log2 (5/14) = 0.940

Notice entropy is 0 if all members of S belong to the same class (the data is perfectly classified). The range of entropy is 0 ("perfectly classified") to 1 ("totally random").

Gain(S, A) is information gain of example set S on attribute A is defined as

Gain(S, A) = Entropy(S) - S ((|S_v| / |S|) * Entropy(S_v))

Where:

S is each value v of all possible values of attribute A

S_v = subset of S for which attribute A has value v

|S_v| = number of elements in S_v

ALGORITHM

1. Decision Tree

2. Tree Generation Algorithm

3. Decision Rules

Decision Tree

With the development of technology, computer, network, database technology is widely used in the daily management. All walks of life have accumulated a wealth of information and data. Database access and query operations cannot meet the requirements. People need to mine more important information from these massive data, such as the overall characteristics of the data description, to discover the interrelatedness of events and predict the development trend of things. Most of the data mining methods use rule discovery or decision tree classification techniques to discover patterns and rules. Its core thought is some kind of induction algorithm. These methods usually mine the data in the database firstly, to generate rules and decision trees, then analyzing the new data and forecasting. The main advantages of these methods are readable by the rules and decision trees. The decision tree classification algorithm in data mining is also an example-based inductive learning algorithm. It looks at a group of no order, no rules, examples, reasoning a decision tree that the formation of the classification rules. The algorithm uses a tree structure to represent a decision set, and using the classification of the data sample set to generate decision rules. Each non-leaf node of the tree represents an attribute test; its branches represent the test results. Each leaf node represents a category. In the decision tree-building process, you need to use pruning to cut the noise in the data and outsiders, thereby improving the reliability of the classification in the unknown data. There are two types of decision trees, classification trees and regression trees. The classification tree is for discrete variables, and regression trees are for continuous variables.

Tree generation algorithm (ID3):

Let S be the set of s data samples, assume that the decision attribute has m different values. Ci(i= 1, ..., m}) is the m different classes, si is the number of samples in the class Ci given by the following formula :

I(s1,s2,…,sm))(log A）

Attribute A has v different values {a1, a2, a3, ..., av }, as the root of the decision tree, S is divided into v subsets {S1, S2, ..., Sv}, which Sj contains some samples of S, and they have the same value of aj in A.

I(s1j,s 2j,…,smj)

The entropy of subsets divided by attribute A is given by equation.

The smaller the entropy value, the higher the purity of the Subsets. The information gain obtained by the branch from attribute A is given by equation (4):

Gain(A)＝I(S1,S2,…,Sm)－H(A)

ID3 algorithm selected attribute A, the largest Gain(A), as the root of the sample set. Subsets of each branch use recursively the ID3 algorithm to build the decision tree nodes and branches, until the sample subsets belong to the same class. This approach makes the minimum average depth and the faster speed of the generated decision tree. A decision tree is generated.

Decision Rules

Decision rules can be extracted from the decision tree. The method is to create a classification rule from the root to the leaf nodes of each path, each attribute - value on the path is the rule antecedent (i.e., IF part) of a conjunctive item. Leaf node is after parts of the rules (i.e., THEN part). These rules can be used to classify new examples. Information assets is a valuable corporate resource, it can exist in various forms of intangible, tangibles, presentation, hardware, software, codes, documentations, tools, services, images etc**.**

Confidentiality, integrity and availability are the three basic attributes in the evaluation of information security. The value of information assets in the risk evaluation is not only the economic value of the assets to measure, but mainly based on the reality and impact of confidentiality, integrity, availability of the information assets. Assets, which property safeties are Different, have different values. Threats to assets, vulnerability and the security measures will impact on asset security attributes. Table 1 is an OA system risk evaluation data table, which assets consist of hardware assets, documents, assets and data assets and system assets.

Characters of ID3 algorithm

Detailed elaborations are presented for the idea on ID3 algorithm of Decision Tree. An improved method called Improved ID3 algorithm that can improve the speed of generation is brought forward owing to the disadvantages of ID3 algorithm. Moreover, based on Improved ID3 algorithm, data mining for Blood-cancers is carried out for primarily predicting the relationship between recurrence and other attributes of breast cancer by making use of SQL Server 2005 Analysis Services. Results prove the effectiveness of Decision Tree in medical data mining which provide physicians with diagnostic assistance. The basic principle of decision tree for constructing tree can be illustrated by ID3 algorithm. It uses the divide-and-conquer strategy in the construction of decision tree, which uses the information gain of characteristic as the heuristic function of attribute selection of a branch in each node of the tree, selecting the information gain as the characteristic of the branch.

ID3 algorithm is described as follows

Let E = D1× D2× ... × Dn be finite-dimensional vector n, where Dj

is a finite set of discrete symbols, E elements e = <v1, v2, ... , vn> is the sample, vj Dj, j = 1, 2, ..., n. Let PE be the positive sample set, NE be the anti-sample set, and the number of samples which are p and n. According to the principle of information theory,

ID3 algorithm is based on two assumptions

(1) In the vector space E, a decision tree classification probability for any sample and the probability for positive sample and anti-sample in E are the same.

(2) The expected bits of information needed for making the correct identification by a decision tree are:

If attribute A is the root of the decision tree, A has n values {u1, u2, ... , un}, which will divide the sample set E into n subsets {E1, E2, ... , En}. Supposing that Eicontains pi positive samples and negative samples, then a subset of the information needed for the Eiis I (pi+ ni), and the expected information needed for the attribute A as the root node.

Therefore, the information gain of classification attribute ofA as the root node is Gain (A) = I (p, n)-E (A). ID3 algorithm selection contributes the greatest attribute of Gain (A) to a branch of the node attributes, and each node of the decision tree is using this principle until the decision tree is completed (each node of the samples belong to the same class or all Category attributes are used up). One advantage of ID3 is its time of tree construction and difficulty of the task (such as the number of sample set samples, the number of attributes for each sample to study the complexity of the concept of the decision tree nodes) are steadily increasing in linear and the computation is relatively small.

VIII. CONCLUSION

Here when it is compared between bayesian principal component analysis and BI-bayesian principal analysis, BI-BPCA is considered as accuracy based and it has maximum efficiency when it is compared to BPCA. DNA chips and other techniques measure the expression level of a large number of genes, perhaps all genes of an organism, within a number of different experimental samples (conditions). The samples may correspond to different time points or different environmental conditions. In other cases, the samples may have come from different organs, from cancerous or healthy tissues, or even from different individuals. Simply visualizing this kind of data, which is widely called gene expression data or simply expression data, is challenging and extracting biologically relevant knowledge is harder still. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column. Each element of this matrix represents the expression level of a gene under a specific condition, and is represented by a real number, which is usually the logarithm of the relative abundance of the mRNA of the gene under the specific condition. Clustering techniques can be used to group either genes or conditions, and, therefore, to pursue directly. Hence bayesian principle is widely used to predict the cancer disease with the help of ID3 algorithm. Finally a graph is generated to show the comparison between BPCA and BI-BPCA.

IX. SCOPE OF PROJECT

The Scope of the project is in future generation. This can lead to a great success in the medical field. It can have the accurate cancer predicted. However, applying clustering algorithms to gene expression data runs into a significant difficulty. Many activation patterns are common to a group of genes only under specific experimental conditions. In fact, our general understanding of cellular processes leads us to expect subsets of genes to be co-regulated and co-expressed only under certain experimental conditions, but to behave almost independently under other conditions. Discovering such local expression patterns may be the key to uncovering many genetic pathways that are not apparent otherwise. It is therefore highly desirable to move beyond the clustering paradigm, and to develop algorithmic approaches capable of discovering local patterns in microarray data.

IX. ACKNOWLEDGMENT

First and foremost we thank the **Almighty** for being our light and for showering his gracious blessings throughout the course of this project. We are grateful to our beloved Principal **Dr. V.SOUNDARARAJAN,** for his ever tireless and relentless support. We extend our sincere thanks to our Head of the Department **Dr. A.JAMEER BASHA,** for his encouragement and inspiration. We are greatly indebted to our Project guide

Ms. M.SANGEETHA Assistant Professor**,** and Project coordinator **Ms. G. LAVANYA** Assistant Professor, Department of Information Technology for their valuable guidance and suggestions in all aspects that aided us to ameliorate our skills. We are thankful to all the **teaching and non teaching staffs of Department** of **Information Technology** and to all those who have directly and indirectly extended their help to us in completing this project work successfully. We extend our sincere thanks to our **family members** and our beloved **friends**, who had been strongly supporting us in all our endeavors.

REFERENCES

[1] Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. Discovering local structure in gene expression data: The order–preserving submatrix problem. In *Proceedings of the 6th International Conference on Computacional Biology (RECOMB'02)*, pages 49–57, 2002.

[2] Pavel Berkhin and Jonathan Becher. Learning simple relations: theory and applications. In *Proceedings of the 2nd SIAM International Conference on Data Mining*, pages 420–436, 2002.

[3] Stanislav Busygin, Gerrit Jacobsen, and Ewald Kramer. Double conjugated clustering applied o leukemia microarray data. In *Proceedings of the 2nd SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data*, 2002.

[4] Andrea Califano, Gustavo Stolovitzky, and Yunai Tu. Analysis of gene expression microarays for phenotype classification. In *Proceedings of the International Conference on Computacional Molecular Biology*, pages 75–85, 2000.

[5] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pages 93–103, 2000.

[6] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Cliffoord Stein. *Introduction to Algorithms*. The MIT Electrical Engineering and Computer Science Series. The MIT Press, 2nd edition, 2001.

[7] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 269–274, 2001.

[8] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretical co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*, pages 89–98, 2003.

[9] D. Duffy and A. Quiroz. A permutation based algorithm for block clustering. *Journal of Classification*, 8:65–91, 1991.

[10] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. In *Proceedings of the Natural Academy of Sciences USA*, pages 12079–12084, 2000.

[11] Dan Gusfield. *Algorithms on strings, trees, and sequences*. Computer Science and Computational Biology Series. Cambridge University Press, 1997.

Ms. M. Sangeetha received the M.E., degree in Computer Science and Engineering from Anna University, Chennai. She is working as an Assistant Professor in Sri Krishna College of Technology, Coimbatore. She is currently a Research Fellow at Anna University, Chennai. Her current Research includes Data Mining, Database Systems and Big Data. She has been active in the areas of databases and information systems, web technology, web mining and web search. She has published 15 papers in National and International Conferences. She is having eleven years of Teaching experience.

Ms. P. Bhuvaneshwari received the M.E., degree in Computer Science and Engineering from Sri Krishna College of Engineering and Technology, Coimbatore. She is working as as Assistant Professor in Sri Krishna College of Technology, Coimbatore. She is currently working in the areas of Big Data, Software Engineering and Data Mining actively.

Ms. A. Sujitha studying final year B.Tech IT in Sri Krishna College of Technology,Coimbatore. she is currently working in the area of Data Mining.

Ms. P. Nandhini studying final year B.Tech IT in Sri Krishna College of Technology,Coimbatore. She is currently working in the area of Data Mining

Ms. C. Gurulakshmi studying final year B.Tech IT in Sri Krishna College of Technology,Coimbatore. She is currently working in the area of Data Mining.