ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

MICRO ARRAY DATA PREDICTION USING BI-CLUSTER BAYESIAN PRINCIPLE

KALAMANI A M.A., M.Phil. Assistant Professor,
UMAMAHESWARI M, Assistant Professor, MCA
PROFESSIONAL GROUP OF INSTITUTIONS

ABSTRACT

Data generated from microarray experiments often suffer from missing values. As most downstream analyses need full matrices as input, these missing values have to be estimated. Bayesian principal component analysis (BPCA) is a well-known microarray missing value estimation method, but its performance is not satisfactory on datasetswith strong local similarity structure. A bicluster-based BPCA (bi-BPCA) method is proposed in this paper to fully exploit local structure of the matrix. In a 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. An automatic parameter learning scheme is also developed to obtain optimal parameters.Experimental results on four real microarray matrices indicate that bi-BPCA obtains the lowest normalized root-mean-square error on 82.14% of all missing rates. Bayesian principal component analysis (BPCA), biclustering, microarray missing value estimation.

INTRODUCTION

ABOUT DATA MINING

Data Mining Concepts

It is a new terminology used in IT field to find some interesting information from a large database. It is a high-level application technique used to present and analyze data for decision-makers.

It is defined in many ways. Some of the definitions are

- It refers to the finding of relevant and useful information from databases.
- It deals with finding of patterns and hidden information from a large database.

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

- It is also known as Knowledge Discovery in Databases (KDD) which is defined as the nontrivial extraction of implicit, previously unknown and potentially useful information from the data.

Data Mining Applications

Data mining was initially successful in marketing. Now, it is used in many areas like wed mining, banking, medical, scientific research etc. It is widely used in Internet to personalize the website in order to provide the necessary information required by the users in a faster way.

Methodology

Attribute Selection in multi array model

How does ID3 decide which attribute multi array model 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 in multiarray

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))

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

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

Applications of ID3 in multi array model:

- Decision Tree
- Tree Generation Algorithm
- Decision Rules

Cancer Overview

Accurately assessing cancer risk in average- and high-risk individuals and determining cancer prognosis in patients are crucial to controlling the suffering and death due to cancer. Cancer prediction models provide an important approach to assessing risk and prognosis by identifying individuals at high risk, facilitating the design and planning of clinical cancer trials, fostering the development of benefit-risk indices, and enabling estimates of the population burden and cost of cancer.

The Epidemiology and Genomics Research Program (EGRP) has compiled the following list of cancer risk prediction models to serve as a research resource for investigators.

- Bladder cancer
- Breast cancer
- Blood Cancer
- Bone Can
- Cervical cancer
- Colorectal cancer
- Stomach cancer
- Liver cancer

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

- Lung cancer
- Melanoma
- Other cancers or multiple sites
- Ovarian cancer
- Pancreatic cancer
- Prostate cancer
- Testicular cancer

Blood Cancer

Cancerous formation can attack any organic system of the human physiology. As part of blood cancer, the rapidly multiplying cancerous cells are found attacking the different aspects of the circulatory system. Besides blood and the lymphatic system; the bone marrow can also be the focus of attack.

Primarily, there are three basic types of blood cancer. Each of the variety may also include several variations, but in general this cancer is categorized into the following kinds

Leukemia- With spurt in the multiplicity of cancerous cells affecting either the marrow or the blood; the ability of the circulatory system to produce blood is severely impaired with.

Lymphoma- The cancerous formation affecting the lymphocytes is referred to as the lymphoma. Lymphocytes are one of the varieties of white blood corpuscles.

Myeloma- As part of Myeloma, the plasma (another variety of WBC) is affected by the cancerous formation.

Characters of ID3 algorithm in multi array model

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

All Rights Reserved © 2016 IJARBEST

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 in multi array model algorithm is described as follows

Let E = D1× D2× ... × Dn be finite-dimensional vector n, where Djis a finite set of discrete symbols, E elements e = <v1, v2, ... ,vn> is the sample, vjDj, 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:

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.

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

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

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.

EXISTING SYSTEM

Existing Method: Decision tree classification algorithm

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. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining. Decision 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. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that 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.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

the most common strategy for learning decision trees from data, but it is not the only strategy. In fact, some approaches have been developed recently allowing tree induction to be performed in a bottom-up fashion.

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. In a decision tree, all paths from the root node to the leaf node proceed by way of conjunction, or AND. In a decision graph, it is possible to use disjunctions (ORs) to join two more paths together using Minimum message length (MML). 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. The general coding scheme results in better predictive accuracy and log-loss probabilistic scoring. In general, decision graphs infer models with fewer leaves than decision trees.

Disadvantages of the Existing system

The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristics such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree.

Decision-tree learners can create over-complex trees that do not generalise well from the training data. (This is known as over fitting) Mechanisms such as pruning are necessary to avoid this problem.

There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems. In such cases, the decision tree becomes prohibitively large. Approaches to solve the problem involve either changing the representation of the problem domain or using learning algorithms based on more expressive representations (such as statistical relational learning or inductive logic programming).

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

For data including categorical variables with different numbers of levels, information gain in decision trees is biased in favour of those attributes with more levels.

PROPOSED SYSTEM

Proposed Method: Decision tree classification method using modified ID3 algorithm.

Cancer is one of the deadliest diseases found 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. 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.

The proposed sample data used by ID3 has certain requirements, which are:

Attribute-value description - the same attributes must describe each example and have a fixed number of values.

Predefined classes - an example's attributes must already be defined, that is, they are not learned by ID3.

Discrete classes - classes must be sharply delineated. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect.

Sufficient examples - since inductive generalization is used (i.e. not provable) there must be enough test cases to distinguish valid patterns from chance occurrences.

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

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

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.

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 specialised 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:**

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

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.

CONCLUSION

Classification is very essential to organise data, retrieve information correctly and swiftly. Implementing Machine learning to classify data is not easy given the huge amount of heterogeneous data that's present in the web. ID3 algorithm depends entirely on the accuracy of the training data set for building its decision trees. The ID3 algorithm learns by supervision. It has to be shown what instances have what results. Due to this ID3 algorithm, it cannot be successfully classify documents in the web. The data in the web is unpredictable, volatile and most of it lacks Meta data.

The way forward for Information Retrieval in the web, in the future opinion would be to advocate the creation of a semantic web where algorithms which are unsupervised and reinforcement learners are used to classify and retrieve data.

Thus the thesis explains the trends, threads and process of the ID3 algorithm which was implemented for finding the missing values and predicting blood cancer disease in a successfully manner.

FUTURE ENHANCEMENT

Inductive learning algorithms have been suggested as alternatives to knowledge acquisition for expert systems. However, the application of machine learning algorithms often involves a number of subsidiary tasks to be performed as well as algorithm execution itself. It is important to help the domain expert manipulate his or her data so they are suitable for a specific algorithm, and subsequently to assess the algorithm results. These activities are often called pre-processing and post processing.

All Rights Reserved © 2016 IJARBEST

The future enhancement discusses issues related to the application of the ID3 algorithm, an important representative of the inductive learning family. A prototype workbench which has been developed to provide an integrated approach to the application of ID3 is presented.

The design rationale and the potential use of the system are justified. Finally, future directions and further enhancements of the workbench are discussed.

- Can implement for web based application
- Handshakes with Inductive learning algorithm
- Improvisations can be done in the performance Evaluation
- Prediction can be done for all kind of diseases
- In case of huge range of data set, data load balancing can be done

REFERENCES:

[1] 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," Nat. Genet., vol. 30, no. 1, pp. 41–47, 2002.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531–537, 1999.

[3] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner, "Multivariate measurement of gene expression relationships," Genomics, vol. 67, no. 2, pp. 201–209, 2000.

[4] R. J¨ornsten,H.Wang,W. J.Welsh, andM.Ouyang, "DNAmicroarray data imputation and significance analysis of differential expression," Bioinformatics, vol. 21, no. 22, pp. 4155–4161, 2005.

[5] L. P. Bras and J. C. Menezes, "Dealing with gene expression missing data," Syst. Biol. (Stevenage), vol. 153, no. 3, pp. 105–119, May 2006.

All Rights Reserved © 2016 IJARBEST

[6] A. W. Liew, N. F. Law, and H. Yan, "Missing value imputation for gene expression data: Computational techniques to recover missing data from available information," Brief Bioinform., vol. 12, no. 5, pp. 498–513, Sep. 2011.

[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520– 525, 2001.

[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520– 525, Jun. 2001.

[9] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," Bioinformatics, vol. 19, no. 16, pp. 2088–2096, Nov. 1, 2003.

[10] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: Local least squares imputation," Bioinformatics, vol. 21, no. 2, pp. 187–198, Jan. 15, 2005.

[11] Z. Cai, M. Heydari, and G. Lin, "Iterated local least squares microarray missing value imputation," J. Bioinform. Comput. Biol., vol. 4, no. 5, pp. 935–957, Oct. 2006.

[12] X. Zhang, X. Song, H.Wang, andH. Zhang, "Sequential local least squares imputation estimating missing value of microarray data," Comput. Biol. Med., vol. 38, no. 10, pp. 1112–1120, 2008.

[13] W. K. Ching, L. Li, N. K. Tsing, C. W. Tai, T. W. Ng, A. Wong, and K. W. Cheng, "A weighted local least squares imputation method for missing value estimation in microarray gene expression data," Int. J. Data Mining Bioinform., vol. 4, no. 3, pp. 331–347, 2010.

[14] K. O. Cheng, N. F. Law, and W. C. Siu, "Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data," Pattern Recog., vol. 45, no. 4, pp. 1281–1289, 2012.

[15] T. H. Bø, B. Dysvik, and I. Jonassen, "LSimpute: Accurate estimation of missing values in microarray data with least squares methods," Nucl. Acids Res., vol. 32, no. 3, pp. e34.1–e34.8, 2004.

[16] M. Ouyang, W. J. Welsh, and P. Georgopoulos, "Gaussian mixture clustering and imputation of microarray data," Bioinformatics, vol. 20, no. 6, pp. 917–923, Apr. 12, 2004.

All Rights Reserved © 2016 IJARBEST

[17] M. K. Choong, M. Charbit, and H. Yan, "Autoregressive-model-based missing value estimation for DNA microarray time series data," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 1, pp. 131–137, Jan. 2009.

[18] R. Jornsten, H. Y. Wang, W. J. Welsh, and M. Ouyang, "DNA microarray data imputation and significance analysis of differential expression," Bioinformatics, vol. 21, no. 22, pp. 4155–4161, 2005.

[19] X. Pan, Y. Tian, Y. Huang, and H. Shen, "Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach," Genomics, vol. 97, no. 5, pp. 257–264, 2011.

[20] X. Gan, A. W. C. Liew, and H. Yan, "Microarray missing data imputation based on a set theoretic framework and biological knowledge," Nucl. Acids Res., vol. 34, no. 5, pp. 1608–1619, 2006.

[21] Y. Cheng and G. M. Church, "Biclustering of expression data," in Proc. Int. Conf. Intell. Syst. Mol. Biol., 2000, pp. 93–103.

[22] X. Gan, A. W. Liew, and H. Yan, "Discovering biclusters in gene expression data based on high-dimensional linear geometries," BMC Bioinform., vol. 9, no. 1, pp. 209–223, 2008.

[23] W. C. Tjhi andL.Chen, "A partitioning based algorithm to fuzzy co-cluster documents and words," Pattern Recog. Lett., vol. 27, no. 3, pp. 151–159, 2006.

[24] S. Das and S. M. Idicula, "Application of cardinality based grasp to the biclustering of gene expression data," Int. J. Comput. Appl., vol. 1, no. 18, pp. 47–54, 2010.
[25] Z. Wang, C. W. Yu, R. C. C. Cheung, and H. Yan, "Hypergraph based geometric biclustering algorithm," Pattern Recog. Lett., vol. 33, no. 12, pp. 1656–1665, 2012.

[26] R. Ji, D. Liu, and Z. Zhou, "A bicluster-based missing value imputation method for gene expression data," J. Comput. Inf. Syst., vol. 7, no. 13, pp. 4810–4818, 2011.

[27] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," Neural Comput., vol. 11, no. 2, pp. 443–482, 1999.

[28] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in 15th. Proc. Conf Uncertainty Artif. Intell, 1999, pp. 21–30.

[29] G. N. Brock, J. R. Shaffer, R. E. Blakesley, M. J. Lotz, and G. C. Tseng, "Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes," BMC Bioinform., vol. 9, no. 1, pp. 12–24, 2008.

All Rights Reserved © 2016 IJARBEST

[30] D. N. Baldwin, V. Vanchinathan, P. O. Brown, and J. A. Theriot, "A geneexpression program reflecting the innate immune response of cultured intestinal epithelial cells to infection by Listeria monocytogenes," Genome Biol., vol. 4, no. 1, pp. R2.1–R2.14, 2003.

[31] M. Ronen and D. Botstein, "Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source," Proc. Natl. Acad. Sci. U.S.A., vol. 103, pp. 389–394, 2006.

[32] N. Ogawa, J. DeRisi, and P. O. Brown, "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis," Mol. Biol. Cell, vol. 11, no. 12, pp. 4309–4321, 2000.

All Rights Reserved © 2016 IJARBEST