K.Lakshmipriya *et al*, International Journal of Computer Science and Mobile Applications,
Vol.5 Issue. 10, October- 2017, pg. 174-180 **ISSN: 2321-8363**
Impact Factor: 4.123

MICROARRAY USING ADVANCED REGRESSION FOR FIXING LOST DATA

K.Lakshmipriya, Dr. R.Manickachezian

Research Scholar, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India

Department of Computer Science, N G M College (Autonomous), Pollachi, Coimbatore-642001, India

Email id: lakshmipriya028@gmail.com, chezian_r@yahoo.co.in

Abstract

Data missing is the more complicated problem in now a day, especially in hospitals. Most of the hospitals are using client server technology for data transferring inside the hospital. While transferring huge database from one location to another location else in case of data migration data loss may occur. At the time some values from the table or from database may disappear. These problems are said to be as data missing. In order to find out the missing values, sometimes prediction may used to fill the data. Prediction should be more accurate. So here we are implementing a multidimensional array model with modified advanced regression. From the data set, an operational database will be created for the cancer patients and a database for normal patients. This database will be unique and different types of sample data are available. The modified advanced regression will compare the existing spatial database with the normal database from the input database. So the result will be obtained from the dataset. Whether the patient will affect from cancer or not, also their infection ration percentage can be find out, along with the missing values in the database during the time of data migration. An improved advanced regression was introduced, to find the missing values in the dataset. BPCA obtains the lowest normalized root-mean-square error on 82.14% of all missing rates. Here we are proposing for 95.79 % of accuracy in the improvised algorithm.

Keywords: Database, Missing data, Micro array, Advanced Regression, BPCA, Spatial database.

1. Introduction

Now-a- days, people are going for this new technique to retrieve data from their databases. It is because the volume of their databases has become larger and larger every day. Normally, query tools are used to retrieve data from the database. [1] But, if the database is larger, then it is difficult to retrieve data in an effective way using query tools. Using data mining techniques, the relevant information can be extracted in an effective manner.[2] It is applied only on specific records or historical data in the database and retrieves some interesting or hidden information from the database. Here a data set is available for cancer patient, that data base having the highly possible rates as well as cancer affected patient details. The reading has been taken from the blood samples of the infected patients. The data set range is about 500 members those who affect from blood cancer.

2. Related Works

Microarray gene expression data generally suffers from missing value problem due to a variety of experimental reasons. Since the missing data points can adversely affect downstream analysis, many algorithms have been proposed to impute missing values [5].

The same in different manner in [6] Gene expression microarray experiments can generate data sets with multiple missing expression values. Unfortunately, many algorithms for gene expression analysis require a complete matrix of gene array values as input. But according to the Bayesian principle [7] Gene expression profile analyses have been used in numerous studies covering a broad range of areas in biology. When unreliable measurements are excluded, missing values are introduced in gene expression profiles. [10] This paper discussed about microarrays has gained widespread uses in biological studies such as cancer classification, cancer prognosis and identifications of cell cycle-regulated genes of yeast because of their large number of genes and small size.

3. Existing Method

In the existing method decision tree classification algorithm has been used. 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 to find the missing data. 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. The below formula shows the Entropy using the frequency table of one attribute

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

This existing system deals with decision trees in data mining based missing data prediction. This is the most common method. Decision tree learning with prediction is also method commonly used in data mining. There is no goal is to create a model that predicts the value of a target variable based on several input variables.

4. Proposed System

Attribute Selection in multi array model

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.

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

5. Algorithm

Now-a- days, people are going for this new technique to retrieve data from their databases. It is because the volume of their databases has become larger and larger very day. [1] Normally, query tools are used to retrieve data from the database. But, if the database is larger, then it is difficult to retrieve data in an effective way using query tools. Using data mining techniques, the relevant information can be extracted in an effective manner. It is applied only on specific records or historical data in the database and retrieves some interesting or hidden information from the database.

5.1 Algorithm Steps

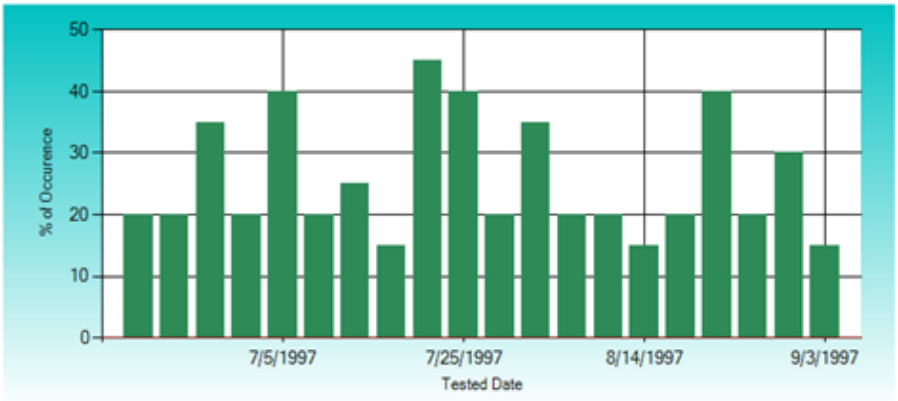
Step 1: Begin the process;
Step 2: Let X be the input dataset (raw data set with missing data);
Step 3: Covert excel into SQL format;
Step 4: String Excel = Server.MapPath("missing data/" + FileUpload1.FileName.ToString()); (Conversion Process)
Step 5 : String connect = "Provider=Microsoft.Jet.OLEDB.4.0;Data Source=" + Excel;
Step 6 : Consider all the data as a data Array;

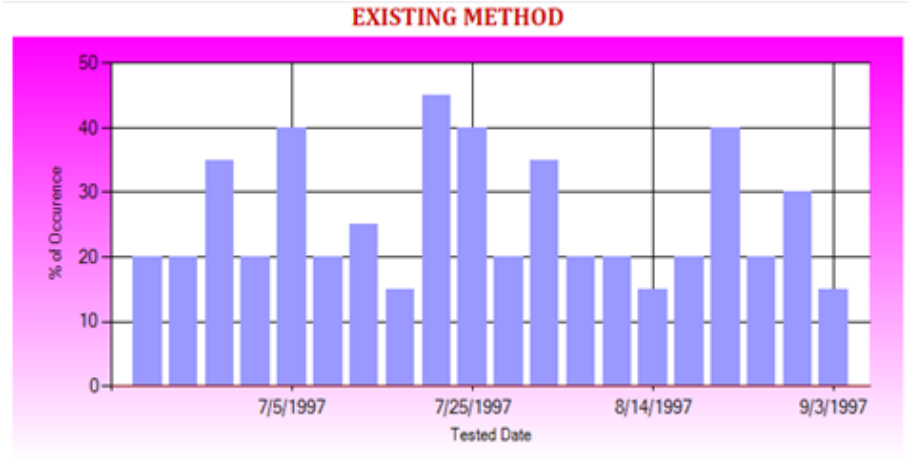
$$= \begin{pmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,p} & A_{1,1} & A_{1,2} & \cdots & A_{1,n-p} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,p} & A_{2,1} & A_{2,2} & \cdots & A_{2,n-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{k,1} & B_{k,2} & \cdots & B_{k,p} & A_{k,1} & A_{k,2} & \cdots & A_{k,n-p} \end{pmatrix}$$

Step 7: B1,1 will be the initial node and Ak,n-p will the end of the row;
Step 8: Array will allow search for row and Column;
Step 9: Finding Null value in the array;
Step 10 : { arr.Add(rd[0].ToString());} : Array execution will done;
Step 11: for (int i = 0; i < n; i++) : Loop will find all missing value for column and row;
Step 12: arr[i].ToString() + : String will verify the value;
Step 13: Let nIG = missing.Count;
Step 14 : if (missing[j] == "") { numofnullIG = numofnullIG + 1;}: finding index for null value
Step 15 : for (int i = 0; i < nIG; i++) : if (missing[i] == "") and inde = i; Fixing X value
Step 16 : val = val + Convert.ToDouble(missing[i].ToString()); Fine tune the result
Step 17 : finalvalue = nIG - numofnullIG; finding min and max value
Step 18 : finalvalue = val / finalvalue;
Step 19 : finalvalue = Math.Round(finalvalue, 2);
Step 20 : Fix value from B1,1 to Ak,n-p will the end of fixing
Step 21 :Repeat the process
Step 22 : Stop

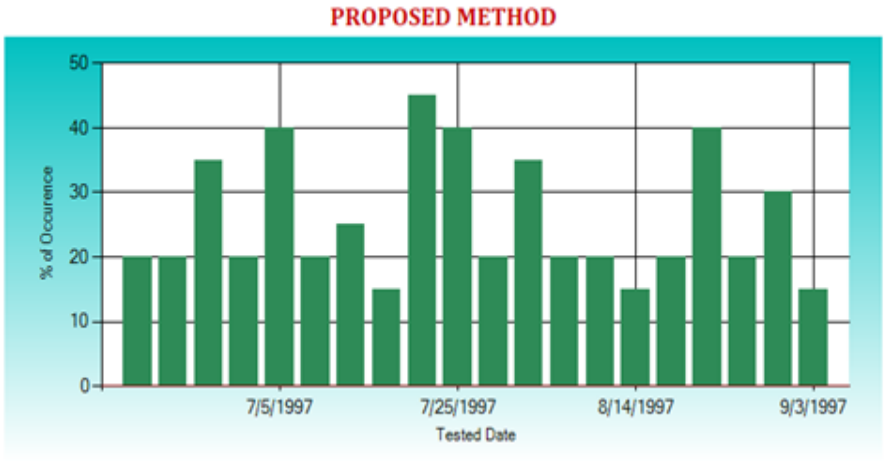
6. Result And Discussion

This is the initial result from the data set. This result shows the occurrence of cancer with data to day percentage calculation of the patient





Tested Date	Percentage of Occurence	CPU Time
6/19/1997	17	8
6/23/1997	17	6
6/27/1997	32	7
7/1/1997	17	6
7/5/1997	37	6
7/9/1997	17	8
7/13/1997	22	7
7/17/1997	12	6
7/21/1997	42	8
7/25/1997	37	9
7/29/1997	17	6



Tested Date	Percentage of Occurence	CPU Time
6/19/1997	20	4
6/23/1997	20	2
6/27/1997	35	2
7/1/1997	20	2
7/5/1997	40	2
7/9/1997	20	4
7/13/1997	25	3
7/17/1997	15	2
7/21/1997	45	4
7/25/1997	40	4
7/29/1997	20	2

From the above given results, the accuracy of the prediction has been improved and the result has been generated in a very short period of time while comparing with the existing system.

7. Future Work

Even our current system works better, we need some enhancement for make the system more better. Our future works related to inductive learning. 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 for a specific algorithm, and subsequently to assess the algorithm results. These activities are often called as post processing and knowledge distribution. The future enhancement discusses issues related to the application of the supervised learning 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 supervised learning 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 to make the system better.

8. Conclusion

Regression 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 from a raw data set. AR algorithm depends entirely on the accuracy of the training data set for building its decision trees. The AR algorithm learns by supervision. It has to be shown what instances have what results. The data in the data set are unpredictable, volatile and most of it contains missing data. The way forward for Information retrieval in the dataset, is considered as a micro array methods according to this algorithms which are unsupervised and reinforcement learners are used to classify and retrieve the missing data. Thus the research explains the trends, threads and process of the AR algorithm which was implemented for finding the missing values and predicting blood cancer disease in a successfully manner.

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, 2002 ,"Introduction to data mining and acpects," Nat. Genet., vol. 30, no. 1, pp. 41–47.

[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, 1999, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531–537.

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

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

[5] A. W. Liew, N. F. Law, and H. Yan , Sep. 2011, "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.

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

[7] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, Nov. 1.2003, "A Bayesian missing value estimation method for gene expression profile data," Bioinformatics, vol. 19, no. 16, pp. 2088–2096.

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

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

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

[11] W. K. Ching, L. Li, N. K. Tsing, C. W. Tai, T. W. Ng, A. Wong, and K. W. Cheng, 2010, "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.

[12] K. O. Cheng, N. F. Law, and W. C. Siu, 2012, "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.

[13] T. H. Bø, B. Dysvik, and I. Jonassen, 2004 ,"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.

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

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

[16] R. Jornsten, H. Y. Wang, W. J. Welsh, and M. Ouyang, 2011, "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.

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

Biography

K.Lakshmipriya received her Bsc (Computer science) from NGM College, Pollachi, India. She completed her Master of Computer Application (MCA) from NGM College, Pollachi, India. Currently, she is a Research Scholar at Department of Computer Science, NGM College, Pollachi, India. She participated in a International Conference. Her area of interest includes Data mining, Missing data.

Dr. R.ManickaChezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published one-fifty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. Recently he received the award "Best Computer Science Faculty of the Year 2015" from Association of Scientists, Developers and Faculties. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

