Estimation of Missing Data Values During Data Migration using AQ Algorithm in Cancer Detection

N. Vikasini[1] Dr. J. Shanthini[2]
[1]PG Scholar [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]SNS College of Technology, Coimbatore, India

Abstract— In Data Mining, data missing is more complicated problem now days, especially in hospitals. Most of the hospitals are using client server technology for data transferring inside the hospital. When migrating database from access to SQL, or Migrating data from SQL to Oracle, data loss may occur. Due to the data loss, some values from the table or from database may disappear. This problem is known as Data Missing. Cancer which is the deadliest disease in which if data is missing including their infection percentage along with the missing values in the database leads to a serious problem. To find out the missing values, sometimes prediction may be used to fill the data. Prediction should me more accurate. So, here we are implementing a multidimensional array model with modified AQ algorithm. An improved AQ algorithm is used to find the missing values with 9.5.79% accuracy in the dataset. 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 AQ algorithm will compare the existing spatial database with the normal database from the input database. So that the result will be obtained from the dataset, whether the patient is affected from cancer or not, including their infection percentage along with the missing values in the database.

Key words: Data Mining, Decision Tree, Missing value, Modified AQ Algorithm, Classification

I. INTRODUCTION

Data mining, the knowledge discovery process analyses data from different perspectives and summarizes it into useful information that increases revenue or cuts costs or both. It extracts knowledge from large databases to discover the existing and newer patterns. The hidden valuable patterns and its relationship are found with automatic technique from huge volume of database which helps in making an effective business decisions. Data mining methods were compared by researchers in many studies that mainly aimed to develop a prediction model in critical fields, like medicine, by investigating several data mining methods, which intended to get the model that have highest prediction accuracy.

Cancer is one of the most common diseases results in majority of death that occurs in any part of the body and spreads to other parts. Detection of cancer at the beginning stage and prevention from spreading to other parts in malignant stage will save the person's life. There are several factors that could affect a person's predisposition for cancer. As data mining technique involves in the use of sophisticated data analysis tools that discovers previously unknown data or pattern, valid patterns and relationships in large data set which can include mathematical algorithm, statistical models, and machine learning methods in early detection of cancer. Four types of learning methods are there, they are classification learning, association learning, and clustering, numeric prediction. In this paper, decision tree is used to classify the data and to mine frequent patterns in data set. The value of the data is tested against a decision tree and a path is traced from root to leaf node which holds the class prediction for that data. Frequent pattern is generated in the dataset with the decision tree. The frequent patterns that are most significantly related to specific cancer types are helpful in predicting the cancer and its type.

Our proposed model reduces the percentage of data missing rate which helps to predict cancer more accurately. So, here we are implementing a multidimensional array model with modified AQ algorithm.

An improved AQ algorithm is used to find the missing values with 9.5.79% accuracy in the dataset. 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 AQ algorithm will compare the existing spatial database with the normal database from the input database. So that the result will be obtained from the dataset, whether the patient is affected from cancer or not, including their infection percentage along with the missing values in the database.

II. LITERATURE SURVEY

In data mining, missing values leads to difficulty in extracting useful information from that dataset [2]. Missing data are the absence of data items in the database that hide some information that may be important [1]. Data mining, the notion method and technique allows to analyses large dataset to extract discover previously unknown structure and relation out of large heap of detail these information is filtered, prepared and classified which will be a valuable aid for decision and strategies [3].

R. Malarvizhi proposed K-Means and KNN methods that provide fast and accurate ways of estimating missing values [5]. According to Y.Fujikawa[6] the phase of handling missing data in KDD process, methods can be classified into two groups: pre-replacing methods and embedded methods. Pre-replacing method deals with missing data in data preparation phase of KDD process and embedded method deal with missing data in data mining phase of KDD process based on this process the former method can be applied more flexibly and the embedded method save more time and cost [7,4]. Edgar Acuna et al [8] prepared a dataset using training sample that shows, the presence of missing values in a dataset can affect the performance of a classifier constructed.

Saleema, J.S [9] presents a paper to find the prominent labels from cancer databases and use the database in a multi-class environment with RAkEL algorithm and

produce a better result than traditional way of predicting cancer.

Title	Author	Method	Advantages	Disadvantages
A Way to Apply Traditional Clustering Methods in Bi-cluster Detection[10]	ZHANG Yanjie, WANG Hong, HU Zhanyi	Bicluster Bayesian component analysis method	1) The missing entries can be easily identified using correlated genes and experimental conditions. 2) This method is better than LLS method	1) Less performance satisfaction on datasets with local similarity structure 2) It Cannot find accurate missed value. 3) BPCA obtains the lowest normalized root-mean-square error on 82.14% of all missing rates
A Comparative Study of Missing Value Estimation Methods[11]	Lim EngAikl, ZaritaZainuddin	Local Least Square imputation method	1) This method allows for the very lage selection of datasets.	1) This method is not consistent in nature.
Application of AQ Algorithm in Information Asset Identification [12]	Sungroh Yoon , Benini.L	AQ algorithm	1) This obtains 95.79% accuracy of all missing rates. 2) It can find accurate missing values.	1) This algorithm is complex in nature

Table 1: Comparisons of prediction algorithm

III. PROPOSED SYSTEM DECISION TREE CLASSIFICATION METHOD USING MODIFIED AQ ALGORITHM

Cancer is a deadliest diseases found among many people across the world. Our project aims to help the medical practitioners to diagnose the patients at the early stage which results in reducing the number of deaths due to cancer. The proposed work is that a modified AQ algorithm using decision tree classification method is developed which 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 which is resolved in pre-processing steps of the data set. And we have proposed an approach to resolve the conflicts in the dataset. Conditional entropy measure concept is used in the AQ algorithm and modified it. The data is processed with pre- processing the data set, and then 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 AQ modified algorithm has certain requirements, which are:

A. Attribute-Value Description

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

B. Predefined Classes

Predefined Classes an example's attributes must already be defined, that is, they are not learned by AQ.

C. Discrete Classes

Discrete 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.

The method called Improved AQ algorithm that can improve the speed of generation is brought forward based on the disadvantages of AQ algorithm. With improved AQ algorithm, data mining for Blood-cancers is carried out for predicting the relationship among recurrence and other attributes of breast cancer by use SQL Server 2005 analysis services. The result proves the effectiveness of using decision tree in medical data mining that provides physicians with diagnostic assistance.

The basic principle of decision tree for constructing tree is illustrated by AQ algorithm. It uses the divide-and-conquer strategy in the construction of decision tree, it 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.

D. AQ in multi array model algorithm

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, 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,

E. AQ algorithm is based on two assumptions

1) In the vector space E, a decision tree classification probability for any sample, 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 of A as the root node is Gain (A) = I (p, n)-E (A). AQ algorithm selection contributes the greatest attribute of Gain

(A) to a branch of the node attributes, and each node in the decision tree is using this principle until the decision tree is. The advantage of AQ algorithm is its time of tree construction and difficulty of the task are steadily increasing in linear and the computation is relatively small.

F. Advantages of Proposed System

- Simple to understand and interpret.
- Requires less data preparation.
- Able to handle both numerical and categorical data.
- Can validate a model using statistical tests.
- Improve the performance.
- Performs well with large datasets in reasonable time.

IV. RESULT ANALYSIS

The proposed system is implemented; the system allows authenticated users to enter the system. A home page will be displayed with current date and current time. Once authenticated user logins, this home page is displayed. In this page upload patient details option is chosen to upload the required patient detail. By choosing the file option, patient report that contains missing data is uploaded. Choose save option to upload the patient report successfully.

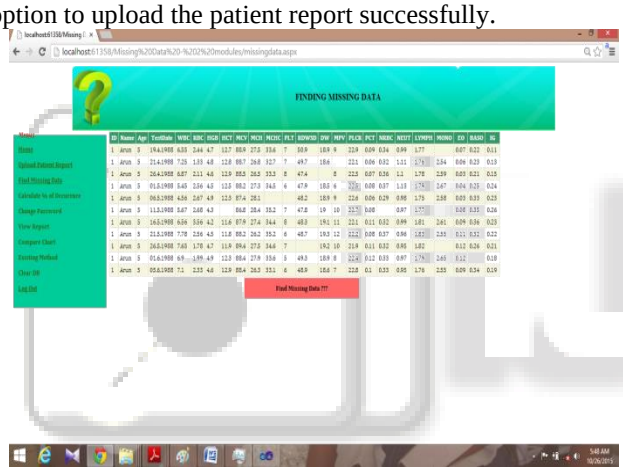

Fig. 1: Uploaded Patient Report

Next choose find missing data button to highlight the missing values in the uploaded patient report. Once find missing data option is chosen, missing data in the uploaded patient report will highlighted. These missing data will be recovered in the future modules. The proposed model is reliable to the existing models.

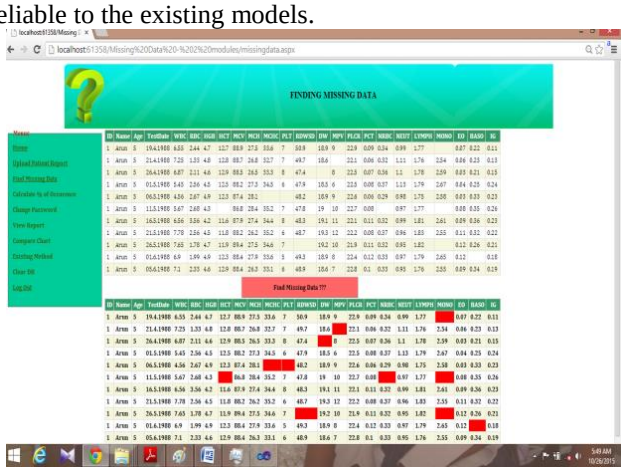

Fig. 2: Finding Missing Data

A. Comparison of LLS, BPCA and RBFIN

Neither the LLS nor BPCA methods required parameter optimization. The performance of each method as a function of the percentage of MVs (2%, 5%, 10% and 15%) for the Lim Yogan, Calen and Oba data sets. The performance is judged by the NRMSE, and tends to decrease with increasing percentage of MVs for each method. The relative performance of the imputation methods did not vary much with the percentage of MVs. It should be noted that in this paper, neither RBF network, BPCA nor LLS required parameter optimization, nor LLS has parameter selection built into the algorithm internally, making these algorithms an attractive choices for automated imputation of MVs.

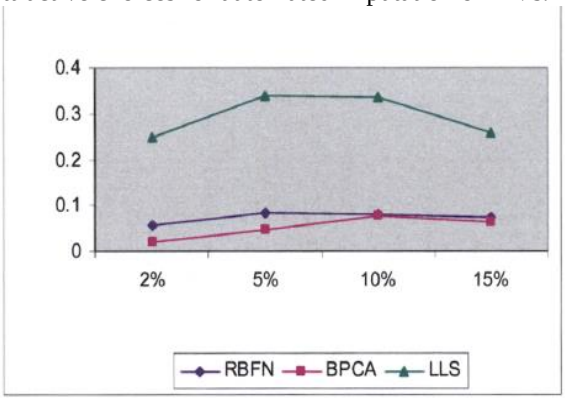

Fig. 3: NRMSE values for different percentage of missing values in Lim & Yogan dataset

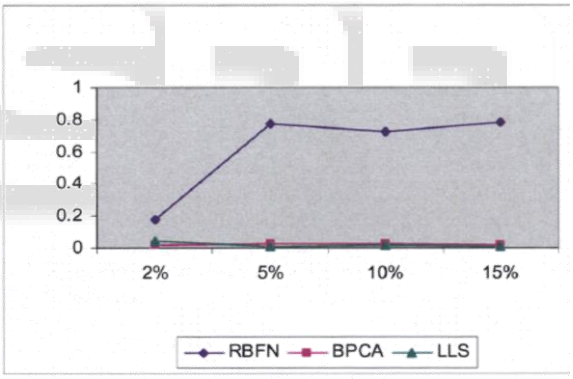

Fig. 4: NRMSE values for different percentage of missing values in calen dataset

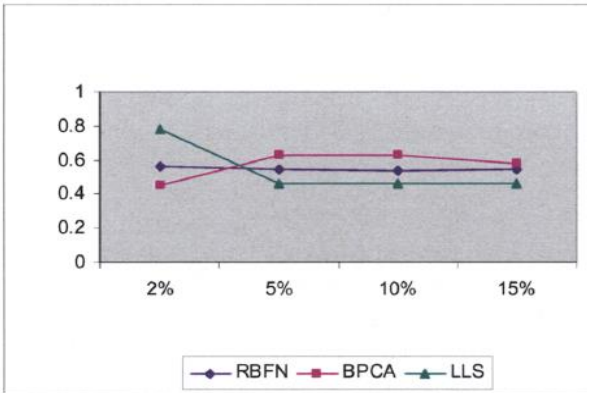

Fig. 5: NRMSE values for different percentage of missing values in obtain dataset

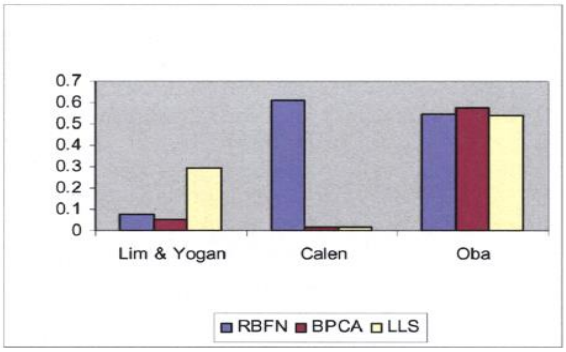

Fig. 6: Average NRMSE values of all missing values for all dataset

We performed an extensive evaluation of existing methods for imputing missing values for data. Here, our investigation demonstrates that the imputation algorithms are highly competitive with each other. The imputation method employed in this paper, RBF network was clearly bested by the more sophisticated methods we tested (LLS and BPCA). Therefore, RBF network cannot be the best choice for imputing MVs. Overall; the BPCA imputation methods performed the best in our simulation study. The LLS algorithm is based on neighbor's selection for imputation, but they also each have features which resembles global based imputation. LLS also allows for the selection of very large number of data .The BPCA is more consistent than the other methods over the data sets we investigated (worst performance of 3rd best in Oba data set, refer to Figure 3), This is due to that the BPCA is less affected by dimension reduction because of the probabilistic model shrinks the principle axes .Therefore, BPCA is quite robust to the changes in data complexity, and its performance is relatively stable over the range of dimension values, and thus has the overall advantage and lowest NRMSE. However, we emphasize that the overall differences between these top methods (LLS and BPCA) is slim and only BPCA method with the lowest overall NRMSE is concluded as the best method.

V. CONCLUSION

Classification is very essential to organize 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. AQ algorithm depends entirely on the accuracy of the training data set for building its decision trees. The AQ algorithm learns by supervision. It has to be shown what instances have what results. Due to this AQ 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 AQ algorithm which was implemented for finding the missing values and predicting blood cancer disease in a successfully manner.

VI. FUTURE ENHANCEMENT

The future enhancement discusses issues related to the application of the AQ algorithm, is an important representative of the inductive learning family. A prototype workbench which has been developed to provide an integrated approach to the application of AQ 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.
- 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] Dinesh J. Prajapati ,Jagruti H. Prajapat, "Handling Missing Values: Application to University Data Set" .Issue 1, Vol.1 (August- 2011), ISSN 2249-6149.
[2] Luai Al Shalabi, Mohannad Najjar and Ahmad Al Kayed, A framework to Deal with Missing Data in Data Sets . Journal of Computer Science 2 (9): 740-745, 2006 ISSN 1549-363.
[3] Johannes Gambier, Andreas Rudolph," Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery", 303–360, 2002.
[4] Liu Peng, Lei Lei , A Review of Missing Data Treatment Methods.
[5] Ms.R.Malarvizhi, Dr.Antony Selvadoss Thanaman, "K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation", IOSR Journal of Computer Engineering.
[6] Yoshikazu Fujikawa, Efficient Algorithms for Dealing with Missing values in Knowledge Discovery, Master Degree Thesis, Japan Advanced Institute of Science and Technology, 2001.
[7] Gustavo E. A. P. A. Batista and Maria Carolina Monard, An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence 17(5-6): 519-533 , 2003.
[8] Edgar Acuna, Caroline Rodriguez, "The treatment of missing values and its effect in the classifier accuracy".
[9] Saleema, J.S. ; Sairam, B. ; Naveen, S.D. ; Yuvaraj , K. ;Patnaik, L.M,"Prominent label identification and multi-label classification for cancer prognosis prediction",. TENCON 2012 - 2012 IEEE Region 10 Conference Digital Object Identifier: 10.1109/TENCON.2012.6412321 Publication Year: 2012 , Page(s): 1 – 6
[10] Zhang Yanjie ; Wang Hong ; Zhanyi Hu," A Way to Apply Traditional Clustering Methods in Bi-Cluster Detection" IEEE Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference, April 2010, Pages 1-4.
[11] Eng Aik Lim, Zarita Zainuddin,"A comparative study of missing value estimation methods",IEEE Electronic Design, 2008. ICED 2008. International Conference, Dec. 2008,Pages:1-5.
[12] Sungroh Yoon, Benini.L, "Application of AQ Algorithm in Information Asset Identification".

