CLASSIFY THE STUDENT WITH MISSING VALUE TO CALCULATE FUTURE SEMESTER RESULT FOR PLACEMENT RECORD USING KNOWLEDGE ACQUISITION

V. Srinivasan
Senior/Lecturer,
Department of MCA,
Velalar College of Engineering and Technology, Thindal,
Erode(Dt) - 638 012
Email:newsrini@rediffmail.com

Dr. G. Rajenderan
Professor & Head,
School of Science & Humanities,
Kongu Engineering College,
Perundurai,
Erode (Dt)– 638 052.
Email:rajendranjv@gmail.com

Ms. J. Vandar Kuzhali
Senior Lecturer,
Department of MCA,
Erode Sengunthar Engineering College, Thudupathi,
Erode (Dt),– 638 057.
Email:vandarkuzhali@yahoo.com

M. Aruna
Lecturer, Department of MCA,
Velalar College of Engineering and Technology,
Thindal, Erode(Dt) - 638 012
Email:saiaruna2005@yahoo.com

ABSTRACT

The classification problem is one of the main issues in data mining because it aims to extract a classifier which can be used to predict the classes of objects whose class table are unknown. In the case of classification with complete data, many algorithms have been presented in literature, In the case of classification with incomplete data very few algorithms are there in the literature The situation when information about the value of some features is unknown is not hypothetical In the first case, the lack of information is due to the impossibility of performing some tests when some data are missing some tests can be unnecessary when the classifier could make certain decisions without these test results. In industrial diagnosis, the classifier can work in the online process monitoring the failure of some measuring elements does not have to cause stopping the process if the values given by the rest of the measuring elements are enough to make a decision such that the process is performing properly. When available information is not sufficient then the classifier algorithm classifies the result wrongly. When this values or missed the other algorithm eliminate these database of record and formulate their result which might not have proper classification. To resolve this problem we introduce knowledge acquisition for classifying the incomplete data and missing data.

Keywords- classification, industrial diagnosis, classifier algorithm, knowledge acquisition

1. INTRODUCTION

The incompleteness of data is often viewed as a negative aspect of real-world databases. One of the convenient solutions to dealing with incomplete data is to eliminate from the data set those records that have missing values [3]. This approach ignores potentially useful knowledge hidden in the incomplete data. The second solution to dealing with missing data is to estimate the missing value in the data item. In general cases, one may use some expected value in the data item based on statistical measures [1]. However, distributions of data in surveys are commonly unknown. Imputation and use of an expected value for a particular missing data item in these cases might be misleading [2]. The third approach to dealing with missing data is to use generic "unknown" for all missing items and treat incomplete data as complete data [2]. However, few studies of this approach rise above making all missing values equivalent. The fourth approach to dealing with missing data is to treat missing data as non-deterministic data. The missing data item is replaced with two data items, and each data item has its value attached with a probability.

The major problem of this approach is the estimation of the probabilities of the values for the missing data. Since the records with missing items and the records without missing items often come from different populations. The above four conventional approaches to dealing with missing data place their focal points on the side of complete data, but fail to take into account the knowledge about the missing data. This study places the focal point on discovering knowledge about missing data. It uses a rough set rule induction approach to discovering knowledge about the missing data. Rough set theory [5] is a promising approach for knowledge discovery from databases. It is used for this study because survey data are discrete, it requires few assumptions more than the data set itself, and the rule induction process is relatively easy to understand. The rest of the paper is organized as follows. First Section proposes a set of prototypes of knowledge pertinent to missing data in the form of association rules. Next Section presents the knowledge discovery results on a real-world student placement record database using the rough set method.

With the development of database and Internet, people have vast quantities of data. How to discover the interesting knowledge from great capacity of data has become a heated topic. Thus, data mining technology has attracted a lot of researchers and developers. Data mining also called knowledge discovery in databases (KDD), is the process of discovering interesting knowledge from a large amount of data stored either in databases,

data warehouse, or other information repositories [6]. There are a lot of functionalities of DM. However, in real-world datasets, there are many problems in data quality such as incompleteness, redundancy, inconsistency, noise data etc. All these serious problems in data quality will affect the performance of DM algorithms and decision-making [7]: Missing data is a quite common problem of data quality in real-world datasets. There are a lot of reasons accounting for missing data. When collecting data, some informants could not or refused to provide information. When inputting data, some operators' mistakes may lead to missing data problem. During experiments, broken-down machines may cause damage to datasets. This paper will study the effect of missing data to classification models by knowledge acquisition. This paper will emphasize the effect of missing data to classification algorithms.

2. DATA QUALITY AND KNOWLEDGE ACQUISITION

a. Data quality

A major problem is that the data in databases or data warehouses are often "dirty". Broadly, duty data can be divided into three kinds: missing data, not missing but wrong data, and not missing and not wrong but unusable data [8]. The unusable data are dirty data that arise due to differences between data maintained in more than two independent databases or due to incomplete or nonstandard specification of data in one database. While, commerce data quality tools today do not address many types of dirty data, and most types of dirty data examinations and repairs are done by humans with domain expertise. Various data mining software provide several tools to deal with missing data. The typical way is to replace missing data with a mean or mode value. Some kinds of wrong data can be prevented by user-specified integrity constraints enforced by database and transaction processing systems. While, most wrong data and not wrong but unusable data can only be repaired or prevented manually or semi-manually by human experts.

b. Knowledge Acquisition

Computer models have been widely applied to solve real-world problems. For example, whether an investment plan is feasible or not; how much to invest; what will impact the investment environment? Generally speaking, models consist of many parameters, whose changes will affect the outputs of models. Uncertainty is an indispensable property of most models. On most cases, the determination of parameters is quite difficult and complex. However, the uncertainty of parameters will reduce the quality guarantee of models and has an adverse impact on the application of models. Knowledge acquisition is to study the impacts of one or more input variables on the outputs of a model, that is, the knowledge of the model to one parameter or a combination of parameters [9]. Knowledge acquisition can help to determine the dependency of the model on the structure and hypothesis of the environment. Therefore, the accuracy of the parameters should be highly guaranteed. Knowledge acquisition can identify the decisive input parameter of the model and enhance the reliability and prediction accuracy of models. Knowledge acquisition can also provide the best output of a model with different combination of parameters.

1) Find the parameters (variables) which have great effect on the outputs of a model;
2) Study the possible changes of those parameters (variables);
3) Determine how those parameters affect the final decision-making ;
4) Identify what activities will lighten the effects.

Knowledge acquisition has been applied in various fields including medicine, risk analysis, environment engineering, economics, *ecology*, statistics and others. Generally, we can be divided into three categories.

1) Mathematical methods assess sensitivity of a mode1 output to the range of variation of an input. They can assess the impact of range of variation in the input variables on the output, *so* they can be help finding the most important inputs. Mathematical methods include nominal range Knowledge acquisition

2) Statistical methods Statistical methods involve conducting simulations in which inputs are assigned probability distributions and assessing the impact of variation in inputs on the output distribution. Depending on the method, one or more inputs are varied at a time. Statistical methods allow identifying the effect of interactions among multiple inputs. Statistical methods include regression analysis, analysis of variance etc.,.

3) Graphical methods give representation of knowledge in the form of graphs or charts. Generally, graphical methods are used to give visual indication of how an output is affected by variation in inputs. Graphical methods can be quite helpful before further analysis of a model or to describe complex dependencies between inputs and outputs.

3. ROUGH SET METHOD

Rough set theory was developed in the early 1980s [10]. It concerns itself with classificatory analysis based on approximation of sets. Rough sets techniques have advantages over statistical methods in that they typically deal with categorical discrete data which do not meet with statistical assumptions. Numerical data must be discredited in order to apply a rough set technique.

In comparison with other association rule induction techniques, rough set techniques require few assumptions more than the data set itself. Here, we briefly overview the rough set rule induction method with emphasis on missing data, and refer to [10] for comprehensive reviews of rough set theory. In our student placement record evaluation survey shown in Table 1. The database is called decision table, or information system. The example decision table has two observations, or objects. Each object is described by three attributes. The last attribute in this table denotes the decision and is called decision attribute. The other attributes are referred to as condition attributes.

Table1. Example of Survey Database

S.No	Result	Evaluation	Decision
1	Good	Continuous	Good
2	Poor	Non continuous	Poor

Decision rules admissible in Table 1 are the following.
if (Result is Good) and (Evaluation is Continuous)
then Decision is Good
with Support 1, Accuracy 1, Coverage 1
if (Result is Poor) and (Evaluation is Non Continuous)
then Decision is Poor
with Support 1, Accuracy 1, Coverage 1

The support of a decision rule is the number of objects that match the rule. Accuracy is defined as the number of objects that match the rule divided by the number of objects that match the condition, and Coverage is defined as the number of objects that match the rule divided by the number of objects that match the decision. Support, accuracy, and coverage are used to assess the creditability of the rule. The general form of decision rules is
if (condition-1) and (condition-2) and ...
then (decision-1) or (decision-2) or ...
with support, accuracy, coverage
Suppose there are missing data in the database, as shown in Table 2. The question mark (?) in Table 2 indicates the missing value. To find the pattern of missing data, we rearrange the table, as shown in Table 3.

Table2. Example of survey database

S.No	Result	Evaluation	Decision
1	Good	Continuous	Good
2	Poor	Non contiguous	Poor
3	?	Continuous	Good

Table3. Example of Survey Database

S.No	Evaluation	Decision	Result
1	Continuous	Good	Good
2	Non contiguous	Poor	Poor
3	Continuous	Good	?

Using the rough set rule induction method, we can have the following rules, if we keep the object with missing data in the table.

if (Evaluation is Continuous) and (Decision is Good)
then (Result is Good) or (Result is ?)
with Support 2(of 3), Accuracy 0.5,0.5, Coverage 1,1
if (Evaluation is Non contiguous) and (Decision is Poor)
then Result is Poor
with Support 1 (of 3), Accuracy 1, Coverage 1

Since these rules do not have a decision context, they are called association rules in this study. We are particularly interested in the first association rule that is related to missing data. Apparently, as long as the last attribute has a missing value, there normally is inconclusive conclusion in rules. Given the fact that there certainly is a missing value in **Result**, we can perceive knowledge about the missing data based on the rule. We rewrite the part of the rule that is specifically relevant to the missing value as follows.
if (Evaluation is contiguous) and (Decision is Good) then (Result is ?)
with Support 1 (of 3), Accuracy 0.5, Coverage 1

An interpretation of this rule is: the value in Result is missing when Evaluation is contiguous and **Decision** is Good, with 0.3 support, 0.5 accuracy, and 1 coverage. Note that such an association rule does not lend itself to estimation of missing values; rather, it reveals knowledge about missing data in relation to other data items of the database.

4. EXPERIMENTAL RESULT

Classification algorithms commonly consist of two steps. The first step is to build up a classification model using training subset. Generally, models are expressed in the form of classifying rules, decision trees and mathematic formulas. The second step **is** to run the model on testing subset to do predictive classification. There into, building up a model **is** a supervised process **of** learning, that is, finding the classification rules and building up prediction models by studying the known data. Therefore, training subsets are crucial for step one. Quality and quantity of the known

data have a great impact **on** the predictive model. Missing data, noise data and other data quality problems will reduce the accuracy of the predictive model and do harm to the application of the model. It is impossible to build a convictive classification model with incompleteness or noise data. This section will study the impact of missing data to classify accurately.

a. Design of experiments

Researchers of machine learning, statistics and neurology have put forward many classification algorithms such as the Bayesian classifier, decision tree, neural networks, linear regression, K-nearest neighbor's classifier, fuzzy sets and fuzzy logic. This paper **has** selected three representative classification models to study **the** impact of missing data to classification models, that is, Back-propagation neural network(BPN), **C4.5** decision-tree **(C4.5)** and Learning vector quantization(LVQ) Among these three models, missing data almost have some impact on BPN but in Decision-tree C4.5 uses a probabilistic approach to handle missing data and other models substitute mean or mode for missing data. All the classification models mentioned in **this** paper are conducted by 'Weka', a data-mining software developed by University of Waikato in New Zealand [11] **.**

Table 1 Sample Datasets

S.No	Datasets	Records	Attri	Classes
1	Student	846	5	2
2	Glass	1888	8	5
3	Diabetes	768	8	3
4	Sonar	208	6	2
5	Shuttle	699	9	2

Five datasets are used in our experiments. All of them are real-world datasets, which come from the Machine Leanings Database Repository at the University **of** California. Table 1 shows the summary of the five datasets. This table gives the classification based on the software waikato where some records are missed due to the missing data the data are not classified correctly which results in the problem of decision making, So to make the correct classification process. We take the missing value and evaluate in the rough set theory and get the missing value through the knowledge acquisition method.

b. Results and analysis

The classification accuracy of three classification algorithms under different missing percentages is displayed in Table 2, which shows the accuracies of all the classification algorithms have an obvious trend of decrease. In general, when the proportion of missing data in the dataset is less than l0%, they have little effect on the classifiers. If the proportion is greater than 20%, the effect should not be neglected. However, by some simple work, like knowledge acquisition for missing data, the negative impact can be reduced significantly. If the proportion exceeds 20%, there is an obvious decrease in the classification accuracy and the missing data should be handled with high cautiousness. Knowledge acquisition method should be chosen to eliminate the negative impact of the missing data and optimize the performance of classifiers. In the real world, there are great quantities of missing data in databases and, usually, the proportion of missing data in the dataset exceeds 20%. Therefore, research on the methods for dealing with missing data is greatly needed since the prediction accuracy of classifiers can be improved and the scope of application can be enlarged.

The more classes it predicts with missing data, the classification accuracy drops faster with the increasing of the missing rate. When the missing rate is up to 2**0%,** the classification accuracy of these datasets drops by nearly 10% at average, almost two times higher than other datasets, Moreover, if a dataset has more attributes but fewer records, the classification accuracy also suffers **a** lot. For example, dataset 'Student' has about 846 records **and 5** attributes; the classification accuracy falls to a great extent.

Table 2 Classification Accuracy for Different Classifiers with and Without Missing Values

Dataset	Applying B/A	Missing Values	BPN	C4.5	LVQ
Student	Before	45%	44.6	59.7	60.8
	After	22%	57.6	62.1	72.8
Glass	Before	58%	82.4	86.3	86.7
	After	26%	85.1	87.9	88.6
Diabetes	Before	30%	65.5	60.4	66.4
	After	15%	68.5	66.5	69.5
Sonar	Before	12%	72.2	75.6	82.2
	After	6%	81.3	80.7	86.5
Shuttle	Before	25%	86.1	85.3	86.1
	After	13%	88	87.5	88.2

From the above table 2 we can see that the percentage of missing values has reduced for the credit from 45% to 22%, Glass from 58% to 26%, Diabetes 30% to 15%, Sonar 12% to 6% and Shuttle 25% to 13%. Also we can come to the conclusion that on applying the Knowledge acquisition method to substitute the missing values we can improve the classification result.

5. CONCLUSION

Missing data may reduce the accuracy of prediction models. This paper mainly studies the impact of missing data to classification algorithms. The Knowledge acquisition method for three representative classifiers to missing data is analyzed in the experiments. The results showed that, with the increasing of the missing rate, the classification accuracies of all the classification algorithms have an obvious trend of decrease. If the proportion of missing data exceeds 20%, there is an obvious decrease in the accuracy of prediction. Methods for missing data treatment should be chosen cautiously to eliminate the negative impact on the classification accuracy and optimize the performance of classifiers.

REFERENCES

1. Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum Likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B39 (1), 1997.

2. Batista G., Monard, M., An analysis of four missing data Treatment methods for supervised learning, Applied Artificial Intelligence, 17(5/6), 2003, 519-533.

3. Little, R. J. A., Rubin, D. B., Statistical Analysis with Missing Data, 2nd Ed. New York: John Wiley and Sons, 2002.

4. Brown, M. L., Kros, J. F., Data mining and the impact of Missing data, Industrial Management & Data Systems, 103(8), 2003.

5. Pawlak, Z., Rough sets and decision analysis, INFOR, 38(3), 2000, 132-144.

6. Hm J. and Kamber M., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2000.

7. Cios K.J. and Kurgan L., 'Trends in Data Mining and Knowledge Discovery". In N.R. Pal, L.C. Jain.

8. Won K. Byoung -JU and etc., "A axonomy of Dirty Data", Data Mining and Knowledge Discovey, 7,2003, pp. 8 1-99.

9. I. *T.* Yao, "Sensitivity Analysis for Data Mining", Proceedings of The 22nd International Conference of NAJTPS, July 24-26. Chicago, USA, 2003.

10. Pawlak, Z., Rough sets, International Journal of Information and Computer Science, 11(5), 1982, 341-356.

BIOGRAPHY

Srinivasan. V received the B.Sc degree in Computer Science from Bharathidasan College of arts and science 1999, MCA degree in Computer Applications from Kongu Engineering College 2002 and M.Phil degree in Computer Science from allagapa University. He is currently working toward PhD degree in Computer Applications Anna University coimbatore. His research interests are Data mining, prediction, classification, pattern recognizing, and database with recent interests focusing on the end use of data mining.

Dr. G. Rajendran received his Ph.D. degree in Mathematics from the Bharathiar University, Coimbatore India in 2004. He is currently working as a Professor and Head in the Department of Mathematics at Kongu Engineering College, Erode, His area of interest are Fuzzy Logic and its appliance in Data Mining. He has organized various conferences. His research interests are Fuzzy Logic, Computer Networks, Duplicate Detection in Large Databases and Image Processing.

Ms. J. Vandar Kuzhali is a Research Scholar doing her research in Kongu Engineering College, Perundurai, Erode, Tamil Nadu, INDIA. She has received B.Sc., M.Sc Degree in Applied Science - Computer Technology. She has completed her M.Phil - Computer Science in 2003 and she has rendered her service in Erode Sengunthar Engineering College, Erode for the past 9 ½ years. Her main research interest includes Early Detection of Cervical Cancer with Fuzzy Data Mining.

Aruna M received the BCA degree in Computer Science from Ponnaiyah Ramajayam College 1999, MCA degree in Computer Applications from P.R Engineering College 2002 and M.Phil degree in Computer Science from annamali University. She is currently as lecturer in Velalar college of engineering and technology. Her area of interests are Data mining and Software engineering.