Volume 2, Issue 4, April 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Discrete and Continuous Missing Values By Using Mixture-Kernel-Based Iteratives

Mr.Shaik Salam
Department of CSE
Sree Vidyanikethan Engineering college

B.Kiran Kumar*
Department of CSE
Sree Vidyanikethan Engineering college
bathalakiran1346@gmail.com

Abstract— **In this paper, a new setting of missing data imputation, i.e., imputing missing data in data sets with heterogeneous attributes referred to as imputing Mixture-Kernel-Based Iteratives data sets. This concept first proposes two consistent estimators for the methods in terms of classification accuracy and root square error at the different ratios by using the non-parametric algorithms.**

Keywords— **data mining, classifications, data sets**

I. INTRODUCTION

Proper handling of missing values is important in all analyses and is critical in some, such as time series analysis. Improper handling of missing values will distort analysis because, until proven otherwise, the researcher must assume that missing cases differ in analytically important ways from cases where values are present. That is, the problem with missing values is not so much reduced sample size as it is the possibility that the remaining data set is biased. The imputation of values where data are missing is an area of statistics which has developed. Several imputation algorithms are now supported by missing values option.

Commonly used methods to impute missing values include parametric and nonparametric regression imputation methods. The parametric method, such as linear regression, is superior while the data set are adequately modeled. However, in real applications, it is often impossible to know the distribution of the data set. Therefore, the parametric estimators can lead to highly bias, and the optimal control factor settings may be miscalculated. For this case, nonparametric imputation method can provide superior fits by capturing the structure of the data set. However, these imputation methods are designed for either continuous or discrete independent attributes. For example, the well-established imputation methods in, are developed for only continuous attributes. And these estimators cannot handle discrete attributes well. Some methods, such as algorithm, association-rule-based method, and rough-set-based method, are designed to deal with only discrete attributes. In these algorithms, continuous attributes are always discretized before imputing. This possibly leads to a loss of useful characteristics of the continuous attributes. There are some conventional imputation approaches, such as , designed for discrete attributes using a "frequency estimator" in which a data set is separated into several subsets or "cells." However, when the number of cells is large, observations in each cell may not be enough to non-parametric estimate the relationship among the continuous attributes in the cell. When facing with mixed independent attributes, some imputation methods take the discrete attributes as continuous ones, or other methods are used. Some reports, for instance, selected to smooth the mixed repressor's, but without taking the selection of bandwidth into account. Therefore, Racine and Li proposed a natural extension of the method in to model the settings of discrete and continuous independent attributes in a fully nonparametric regression framework.

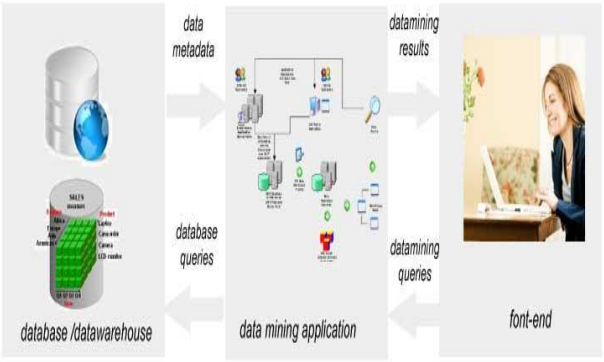
II. RELATED WORK

Data is collected explosively every minute through business transactions and stored in relational database systems. In order to provide insight about the business processes, data warehouse systems have been built to provide analytical reports for business users to make decisions. Data is now stored in database and/or data warehouse system so data mining system should be designed to decouple or couple with these systems. This question leads to four possible architectures of a data mining.

Data mining application layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms. A missing value can signify a number of different things in your data. Perhaps the field was not applicable, the event did not happen, or the data was not available. It could be that the person who entered the data did not know the right value, or did not care if a field was not filled in. Therefore, Analysis Services provides two

distinctly different mechanisms for managing and calculating these missing values, also known as null values.

If the task that you are modeling specifies that a column must never have missing values, you should use the NOT_NULL modeling flag when you define the mining structure. This will ensure that processing will fail if a case does not have an appropriate value. If an error occurs when processing a model, you can then log the error and take steps to correct the data that is supplied to the model. There are a variety of tools that you can use to infer and fill in appropriate values, such as the Lookup transformation or the Data Profiler task in SQL Server Integration Services, or the Fill By Example tool provided in the Data Mining Add-Ins for Excel. However, there are also many data mining scenarios in which missing values provide important information. Generally, Analysis Services treats missing values as informative and adjusts the probabilities to incorporate the missing values into its calculations. By doing so, you can ensure that models are balanced and do not weight existing cases too heavily. This section explains how values are defined and counted as Missing in models that permit null values. This topic also describes how data mining algorithms process and use these Missing values when creating a model.

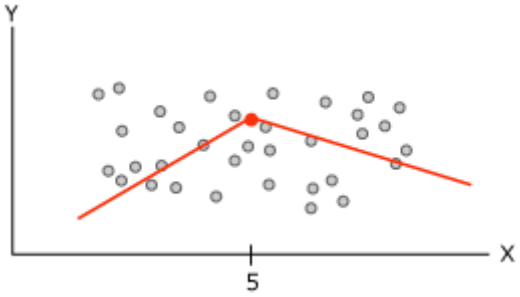


Including the Missing state by default makes sense when you consider that your data might not have examples of all possible values, and you would not want the model to exclude the possibility just because there was no example in the data. For example, if sales data for a store showed that all customers who purchased a certain product happened to be women, you would not want to create a model that predicts that only women could purchase the product. Instead, Analysis Services adds a placeholder for the extra unknown value, called Missing, as a way of accommodating possible other states.

III. NONPARAMETRIC ITERATIVE IMPUTATIONMETHOD

3.1 continuous attribute data sets:

When the Microsoft Decision Trees algorithm builds a tree based on a continuous predictable column, each node contains a regression formula. A split occurs at a point of non-linearity in the regression formula. For example, consider the following diagram.



The diagram contains data that can be modeled either by using a single line or by using two connected lines. However, a single line would do a poor job of representing the data. Instead, if you use two lines, the model will do a much better job of approximating the data. The point where the two lines come together is the point of non-linearity, and is the point where a node in a decision tree model would split. For example, the node that corresponds to the point of non-linearity in the previous graph could be represented by the following diagram. The two equations represent the regression equations for the two lines.

3.2 Discrete attribute data sets:

A set of data is said to be discrete if the values / observations belonging to it are distinct and separate, i.e. they can be counted (1,2,3,....)

When the values in the batch are whole numbers (**counts**), the data set is called **discrete**. Examples of discrete measurements are:
The number of admissions in a hospital's accident and emergency unit each day over a period of two months.
.

Distribution displays for categories are provided which illuminate the distribution of continuous attributes over all cases in a category, and which provide a histogram of the population of the different states of categorical attributes. An array of such displays by attribute (in one dimension) and category (in another dimension) may be provided. Category diagram displays are also provided for visualizing the different categories, and their distributions, populations, and similarities. These are displayed through different shading of nodes and edges representing categories and the relationship between two categories, and through proximity of nodes.

A type of data is **discrete** if there are only a finite number of values possible or if there is a space on the number line between each 2 possible values.

Discrete data usually occurs in a case where there are only a certain number of values, or when we are counting something. Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes. Statistically speaking, discrete data result from either a finite or a countable infinity of possible options for the values present in a given discrete data set. The values of this data type can constitute a sequence of isolated or separated points on the real number line. Each observation of this data type can therefore take on a value from a discrete list of options.

The discrete data type usually represents a count of something. Some examples of this type include the number of cars per family, a student's height, the number of times a person yawns during a day, a number of defective light bulbs on a production line, and a number of tosses of a coin before a head appears.

IV. CONCLUSIONS

A new setting of missing data imputation, i.e., imputing missing data in data sets with heterogeneous attributes referred to as imputing Mixture-Kernel-Based Iterative data sets. This concept first proposes two consistent estimators for the methods in terms of classification accuracy and root square error at the different ratios by using the non-parametric algorithms.

V. FUTERWORK

In future we plan to further explore parametric and non-parametric functions, instead of existing ones, in order to achieve better extrapolation and interpolation abilities in learning algorithms.

REFERENCES

[1] J. Barnard and D. Rubin, "Small-Sample Degrees of Freedom with Multiple Imputation," Biometrika, vol. 86, pp. 948-955, 1999.

[2] G. Batista and M. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," Applied Artificial Intelligence, vol. 17, pp. 519-533, 2003.

[3] H. Bierens, "Uniform Consistency of Kernel Estimators of a Regression Function under Generalized Conditions," J. Am. Statistical Assoc., vol. 78, pp. 699-707, 1983.

[4] C. Blake and C. Merz UCI Repository of Machine Learning Database, http://www.ics.uci.edu/~mlearn/MLResoesitory. html, 1998.

[5] M.L. Brown, "Data Mining and the Impact of Missing Data," Industrial Management and Data Systems, vol. 103, no. 8, pp. 611- 621, 2003.

[6] R. Caruana, "A Non-Parametric EM-Style Algorithm for Imputing Missing Value," Artificial Intelligence and Statistics, Jan. 2001.

[7] K. Cios and L. Kurgan, "Knowledge Discovery in Advanced Information Systems," Trends in Data Mining and Knowledge Discovery, N. Pal, L. Jain, and N. Teoderesku, eds., Springer, 2002.

[8] M.A. Delgado and J. Mora, "Nonparametric and Semi-Parametric Estimation with Discrete Regressors," conometrica, vol. 63, pp. 1477-1484, 1995.

[9] A. Dempster and D. Rubin, Incomplete Data in Sample Surveys: Theory and Bibliography, W.G. Madow, I. Olkin, and D. Rubin, eds., vol. 2, pp. 3-10, Academic Press, 1983.

