Efficient Algorithms for Dealing with Missing Values in Knowledge Discovery

Yoshikazu Fujikawa

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2001

Keywords: efficient algorithm, missing values, large data sets, mean and mode method, cluster-based filling up.

1. Background

Knowledge Discovery in Databases (KDD) has become a new area that attracts not only theoretical researchers but also practitioners. The current progress of processing ability and memory capacity of computers enable us to deal with more data. However, the existence of missing values on datasets is a critical problem of KDD. Data mining methods to extract useful knowledge from datasets have been developed but many of them were not designed for dealing with missing values. Having efficient methods to fill up missing values will extend the applicability of many data mining methods. Also, various methods to deal with missing values have been developed but it is still not clear how different methods can be appropriately used in a given application.

2. Objectives

This research has two objectives:

1. to investigate several well-known methods for processing missing values in KDD by theoretical and experimental comparative evaluation;
2. to find efficient algorithms to deal with missing values for large datasets in data mining. We develop new algorithms to deal with missing values efficiently. The key idea of these

algorithms is they divide a data set that has missing values into some clusters beforehand and replace missing values in each cluster by its mean value or mode value for each numeric or symbolic attribute, respectively.

3. Approach

To investigate well-known methods of dealing with missing values we carry out a survey of existing methods in the machine learning and KDD literature. We propose a classification of missing value occurrence in datasets into three cases in practice that are the occurrence of missing values focusing on some attributes, on some instances, and randomly on different attributes and instances. From the literature we select and evaluate theoretically and experimentally several well-known existing methods. For theoretical evaluation, we investigate eleven methods, mean and mode method, linear regression, the method using standard deviation, nearest neighbor estimator, decision tree imputation, autoassociative neural network, casewise delation, lazy decision tree, dynamic path generation, C4.5 and CART. For experimental evaluation, we implement and compare five methods. The mean and mode method, the method using standard deviation, nearest neighbor estimator, decision tree imputation and C4.5. We estimate error rates of each method by C4.5 classification program after filling up all missing values by using those methods with datasets that were pulled out several values at given rates. From the result we found that, although performance of each method may depend on the type of attribute including missing values on data sets, the mean and mode method often gives high replacing quality in prediction for the datasets.

Based on these results, we propose new algorithms. In order to find efficient algorithms, our main idea is to do cluster-based filling up by mean and mode. We propose three algorithms called Natural Cluster Based Mean and Mode, Attribute Rank Cluster Based Mean and Mode, and k-Means Clustering based Mean and Mode. We do experiments with datasets that have missing values: naturally by the same way as previously. The three algorithms newly proposed are compared to four methods, nearest neighbor estimator, autoassociative neural network, decision tree imputation and C4.5 in terms of error rates and time for replacing. Finally, we try to apply our method to a large dataset, the-census-income dataset. In this trial, we use See5 classification program for estimating error rate of each algorithm. The datasets used in all experiments are from UCI repository. We implement the methods by the C language using UNIX except C4.5 and See5 classification programs.

4. Result

Compared with the other methods, the accuracies in prediction/classification resulted by method that we proposed were the same, or beyond the other ones. Our algorithms could be applied to a dataset which has 300 thousands of instances.

The advantages of our method are:

1. it can deal with missing values with low cost and as accurate for prediction/classification as other methods so that it can be applied to large datasets,
2. it is not a build in method for data mining methods, so that it can be applied to many other data mining methods,
3. it can be applied to unsupervised dataset because it does not require to use the class attribute for its processing.

5. Future work

The following topics are worth to do in the future:

1. though the obtained results are interesting and encouraging, this research is hopefully to be pursued, refined and verified with a larger number of datasets and maybe with other methods of evaluation,
2. we must check effective methods for each missing values cases,
3. we must take into consideration the reason of causing missing values. The cause of missing values may affect the performance of methods that process missing values,
4. if we can measure the relationship between numeric and symbolic attributes, more effective methods to deal with missing values would be developed.