International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Review Article

A Survey and Comparative Study of Different Data Mining Techniques to Implement a Missing Value Estimator System

Ajay Singh Mavai[Å*] and Sadhna K. Mishra[Å]

[Å]Department of Computer Science & Engineering, LNCT Bhopal (M.P.), India

Accepted 15 July 2014, Available online 01 Aug 2014, **Vol.4, No.4 (Aug 2014)**

Abstract

Social media is very useful medium for data collection and predicting status and situation of a user. Many real world applications and data sets often contain many missing elements which may present a big obstacle for many learning algorithms, which usually require a complete data set to build the model. Most algorithms that automatically develop a rule based model are not well suited to deal with incomplete data, till now many missing value estimator system implemented using many techniques such as Bootstrap, K- nearest neighbor, Support Vector Machine, Bayesian Classification, Rough set approach and by using combination of different classifiers, hybrid approach as well as layered approach. In this paper we describe the comparative study of different Data Mining techniques to implement a Missing Value Estimator System (MVES) and also describe the techniques and methods for data selection, finding missing values and modeling to sort the useful information which could be used in other application in different ways.

Keyword: *Missing Value Estimator System (MVES), Social Media(SM), Data Mining, KNN, Bootstrap, SVM.*

Introduction

There are several drawbacks in pattern classification like missing or incomplete data. For instance, wireless sensor networks go through incomplete data due to different reasons, like power outage at the sensor node, random incidences of local interference or a higher bit error rate of the wireless radio transmissions similarly in social media also (Almeida R.J *et al*,2010). Generally, pattern classification with missing data concerns two different problems, managing missing values and sample classification. Depending on how both difficulties are resolved, machine learning approaches can be grouped into four categories. One is termed as ‘‘complete case analysis’’, which disregards the observations with missing values and the study is based on the entire data. This technique limits to only where bulk amount of data is available. The main short comes of this procedure is the loss of efficiency because of removing the incomplete observations. The second technique is the imputation method for handling missing values. Imputation is a set of procedures that want to fill in the missing values with the approximated ones. The main idea behind this technique is to utilize known relationships among the entire values of the dataset to support in missing data evaluation. The third technique is to presume some models for the input data and then a maximum likelihood method find estimates for the models. Missing values for a model can be handling using expectation-maximization (EM) algorithm through the Maximum likelihood method. After creating a model, the classification for a particular input section is execute using the Bayes theorem. The last one technique deal with missing values during the classification process without any imputation with the help of machine learning methods like decision trees and fuzzy neural network. There are several machine learning technique, like standard feed-forward neural networks (FFNN) or support vector machines (SVM), require a complete input data matrix (S. Sumathi *et al*, 2006).

This paper is paying attention on comparative study on missing data estimator system which is resolved through the different techniques of Data Mining. But some of the challenges are arises during building the missing value estimator system are behavior of users, no. of items and relation between users. Estimator systems will have to deal with all of these issues in order to recommend the missing value in given information system. The complexity and granularity of these problems differ, but each represents a real-life area where we believe improvements can be made (Jiawei Han *et al*, 2011). In developing the proposed system, the following practical problems arise:

- To process data defined with according to the technique
- To define the behavior of data.
- To define membership function according our sample information system.

Concept of Data Mining

Data Mining is a relatively new and young interdisciplinary field of computer science, with the main

*Corresponding author:**Ajay Singh Mavai**, **Dr. Sadhna K. Mishra** is working as Professor

purpose is to gather intelligence from data sets which are large in number. Major chunk of the work involves transforming the intelligence gathered from large data sets into human- understandable form. Data Mining is a powerful new technology, used for extraction of hidden predictive information from large databases. Initially, statistical analysis was a viable solution on a collected data manually. However when electronically stored data replaced the collection of data in huge amount, this increased the demand of a useful technique for analysis which works in an automated way. Thus Data Mining is a relatively young and interdisciplinary field of computer science for discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The main objective of Data Mining is to extract knowledge from large data sets in a human understandable structure. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies. So we can say that Data Mining is the Non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data and Exploration and Analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns or rules.

Data mining is the process of discovering meaningful, new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques. Basically, data mining is only one of the many steps involved in knowledge discovery in database and Knowledge Discovery in Database was introduced in 1989 with the concept of being wide and high level in the quest of knowledge from data. Data mining is a step in the Knowledge Discovery in Database process concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations. Hence, the arrangement that are the result of the data mining process must meet certain conditions so that these can be measured as knowledge which are legitimacy, understandability, usefulness, uniqueness and interestingness.

The steps of Knowledge Discovery in Database that are begin with the raw information and concluding with the extracted information are

- Selection or segmentation of the information that are appropriate to some criteria,
- Data cleaning and preprocessing of data where unnecessary information is removed,
- Data transformation and reduction,
- Data mining that concerned with the extraction of patterns from the data,
- Elucidation and Estimate the information which is used to support decision making.

There are two fundamental goals of data mining that are prediction and explanation where prediction makes use of existing variables in the database in order to forecast or predict the unknown values of interest and explanation focuses on finding patterns describing the data and the consequent presentation for user analysis. There are some several data mining techniques that fulfill these objectives are Neural networks, Decision trees, Genetic Algorithms, Nearest neighbor method, Rule Induction etc.

Data Mining Applications: A wide range of companies have deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships (S. Sumathi *et al*, 2006). Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied. Some successful application areas include:

- Games
- Business and e-commerce data
- Science and Engineering
- Human Rights
- Spatial Data Mining
- Web Mining
- Sensor Data Mining
- Visual Data Mining
- Observation
- Pattern Mining
- Subject- based Data Mining
- Health Care Data etc.

Different techniques to implement missing value estimator system

Missing value estimator system can be implemented using different techniques like data mining classification technique, neural network techniques, artificial intelligence techniques etc using different algorithm or combination of two or more algorithm such as decision trees , Bayes classifiers, k- nearest neighbour classifier, case based reasoning, fuzzy logic technique, clustering algorithms, genetic algorithm etc, Several types of algorithms are particularly significant for implementation of missing value estimator system, here we compare some of these algorithm that are used to implement a estimator system such as Bootstrap, is a method for evaluating any statistical procedure and is the part of BOAT stands for Bootstrap Optimistic Algorithm for Tree construction which is approximation algorithm based on sampling, it is not based on the use of any special data structures, as an alternative, it uses a statistical technique known as "bootstrapping" to create large sample in to several smaller samples or subsets of the given training data set. Bootstrap method work on tree pattern which is useful when our search is located on root node. BOAT is usually to be found that two or three times faster than other algorithm like Rain Forest whereas construction of tree are almost exactly same (Zoubir, A.M *et al*, 1998). Moreover the main advantage of BOAT is that it can be used for incremental updates.ie. BOAT can acquire new insertions and deletions for the training data and update the decision tree without reconstruction of the tree. The bootstrap

Table 1 Comparison among Data mining Techniques

Classifier	Technique used	Criterion	Pros	Cons
Rough Set	Rough set theory is based on the establishment of equivalence classes within the given training data that classified to discover structural relationships within Imprecise or noisy data. It applies to discrete-valued attributes. Continuous-valued attributes must therefore be discretized before its use.	A pair of precise sets called the lower and the upper approximation of the rough set is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possible belong to the set.	1-Provides efficient algorithms for finding hidden patterns in data. 2-Identifies relationships that would not be found using statistical methods. 3-Allows both qualitative and quantitative data.	It generates too many rules that create many difficulties while taking decisions
Support Vector Machine	SVM is an algorithm for the classification of both linear and nonlinear data. It transforms the original data in a higher dimension, from where it can find a hyperplane for separation of the data using essential training tuples called support vectors.	Its effectiveness lies in the choice of root and soft boundary parameters. For roots, different couple of (C, γ) values are tried and the one with the most excellent cross-validation accuracy is chosen. Trying exponentially increasing sequences of C is a practical technique to identify good parameters.	1.Highly precise 2.Able to model complex nonlinear decision boundaries 3.Less prone to over fitting than other techniques. 4.It can be used for prediction as well as classification.	1.Suffer from slow processing when training with a large set of data tuples, due to high algorithmic complexity. 2. The selection of the root is not easy. 3.The velocity both in training and testing is time-consuming.
KNN	KNN method selects its K closest observations (neighbors) according to a distance metric.	Selected observations present known values on the features to be imputed. A weighted average of these values is then used as an estimate for each incomplete feature value.	Its selecting only some particular area not taking whole area because of that we can customize our search in particular area.	1. It is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. 2. It must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training sets.
Bayesian Method	On the basis of rule, the combined probabilities of model interpretation and classes, the algorithm tries to approximate the restricted probabilities of classes given an observation.	In Bayes, all model constraints (i.e., class priors and characteristic probability distributions) can be approximated with comparative frequencies from the training set.	1. Naive Bayesian classifier simplifies the computation. 2.It Reveal high precision and velocity when applied to huge amount of databases.	1 The suppositions made in class situational independence. 2. Lack of available probability data.
Decision Tree	BOAT is based on a well-known statistical principal called bootstrapping.	It is a decision tree algorithm to create several smaller subsets of the given training data, each of which fits in memory. Each subset is used to construct a tree, resulting in several trees. The trees are examined and used to construct a new tree, that turns out to be very close to the tree that would have been generated if all of the original training data had fit in memory.	1. It is two to three times faster than RainForest, while constructing exactly the same tree. 2. It can acquire new insertions and deletions for the training data and update the decision tree without reconstruction of the tree. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	1. The bootstrap method works well with small data sets and not appropriate for large data sets. 2. Decision tree algorithms are unstable. 3. It does not provide general finite-sample guarantees.

method samples the given training data regularly with substitution i.e. training data update incrementally, each time a data is selected equally probable to be selected again and readied to the training set. For instance, imagine a social media site that randomly selects users as a data for our training set. In the process of sampling with substitution, the machine is allowed to go for the same user more than once. If we talk about the size of data sample, so according to *Robert Stine [ICPSR Blalock Lectures, 2003]* is that higher quantity of data is always better when it comes to sample size. If the size of a sample is small though, the theory shows that the bootstrap makes better use of the data than traditional methods. Recently researching is using bootstrap very frequently for to train

the model with the data from multi media. The technique is also subcategorizing into several sections, he most common and adaptive is .632 bootstrap. The overall accuracy of the model is more important and essential part of this estimation process that are achieved by repeating the sampling procedure n times, where as in each iteration we use the existing test set to create an accurate estimation of the model acquired from the current bootstrap sample. The accuracy of the model can be estimated using equation _

$$Acc\ (M) = \sum_{i=1}^{n}(0{:}632\ \mathrm{X}\ Acc(\mathrm{Mi})_{\mathrm{test_{set}}}\ set\ + 0{:}368\ \mathrm{x}\ Acc(Mi)_{\mathrm{train\_set}}\ );$$

where *Acc*(*Mi*)*test set* is the accuracy of the model obtained with bootstrap sample *I* when it is applied to test set *i*. *Acc*(*Mi*)*train set* is the accuracy of the model obtained with bootstrap sample *i* when it is applied to the original set of data tuples. The bootstrap method works well with small data sets.

Another approach is the KNN stands for K nearest neighbors (KNN Algorithm) that is one of the most popular methods for solving incomplete data problems. The *k*-nearest-neighbor method was first described in the early 1950s. KNN method selects its K closest observations (neighbors) according to a distance metric, where the selected observations present known values on the features to be imputed. A weighted average of these values is then used as an estimate for each incomplete feature value, the main advantage of this method is its selecting only some particular area not taking whole area because of that we can customize our search in particular area. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition. Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.

Next one is Support Vector Machine that is we compare in this paper, SVM is based on statistical learning theory and is increasingly becoming useful in data mining.SVM is the method that is used for the classification of both linear and nonlinear data (E. Osuna *et al*, 1997). SVM is an algorithm that uses a nonlinear mapping to transform the original training data into a

higher dimension (J. C. Platt *et al*, 2008). Within this new dimension, it searches for the linear optimal separating hyperplane that shows a decision boundary separating the tuples of one class from another class with an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors and margins (Manikandan, J *et al* 2008).

Next approach is Bayesian classifiers, are statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers (Jufen Zhang *et al*, 2004). Bayesian classifiers have also exhibited high precision and pace when applied to large databases. Naïve Bayesian classifiers suppose that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve." Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes (Bose, S. *et al*, 2012). Bayesian belief networks can also be used for classification, and last one approach that we compare here is rough set approach, has recently become a popular theory in the field of data mining introduced by Pawlak in the early 1980s provides a formal framework for the automated transformation of data into knowledge (R.J.A. Little *et al*, 1987, D.B. Rubin.1976). A rough set is a pair of approximation set i.e. lower approximation set that also said to be a positive cases and upper approximation s*et al*so termed as possible cases (Zadeh L A. 1965). Using this approximation, the rough set theory develops tools to discover rules from the given databases. It is tool for handling vagueness and uncertainty inherent to decision situations (Pawlak Z. 1982, Pawlak Z *et al*, 2007). It simplifies the search for dominating attributes leading to specific properties, or just rules pending the data. Though rough set has major advantages over the other methods, but it generates too many rules that create many difficulties while taking decisions (Pawlak Z, *et al*, 2007). Therefore, it is essential to minimize the decision rules.

Conclusion

This paper has offered a survey and comparative study of the various data mining techniques that have been projected in the direction of improvement to implement Missing value Estimator Systems. Many authors have come up with a number of algorithms that can be used to mine the attitude of real time application like online users of the SM. We have shown the ways in which data mining has been known to facilitate the development of Estimator system and the ways, in which the various techniques have been applied, we compared the above following method and projected to contribute in the effort of researchers to build a better and effective model to discover missing value. The consequences of this study are estimated to support different entities to get back essential information on a particular problem and consequently using this information as a decision to suggest missing values.

References

Almeida R.J, Kaymak U. ; Sousa, (2010), A new approach to dealing with missing values in data-driven fuzzy modeling, *Fuzzy Systems (FUZZ),IEEE International Conference* pg 1-7.

S. Sumathi, S.N. Sivanandam (2006) Introduction to Data Mining and its Applications, *Studies in Computational Intelligence*, Volume 29, Springer 828.

Jiawei Han and Micheline Kamber (2011) Data Mining: Concepts and Techniques, *Second Edition Elsevier*- 744.

Zoubir, A.M. (1998) The bootstrap and its application in signal processing, *Signal Processing Magazine, IEEE* (Volume: 15, Issue: 1) pg. 56-76.

E. Osuna, R. Freund, and E Girosi (1997) Improved Training Algorithm for Support Vector Machines, *Neural Networks for Signal Processing, IEEE Workshop* Page(s): 276 – 285.

J. C. Platt, Scholkopf, Burges, and Smola,(2008) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Advances in Kernel Method: *Support Vector Learning*, by MIT Press, pg. 185- 208.

Manikandan, J. ; Venkataramani B.(2008) Diminishing learning based SVM classifier with non-linear kernels *IEEE International Conference on Electronic Design (ICED 2008) pg*. 1-6.

Jufen Zhang ; Everson, R.(2004) Bayesian estimation and classification with incomplete data using mixture models *Machine Learning and Applications. IEEE International Conference on Machine Learning and Applications*, pg. 296-303.

Bose, S. ; Das, C. ; Dutta, S. ; Chattopadhyay, S.(2012) A novel interpolation based missing value estimation method to predict missing values in microarray gene expression data**:** *Communications Devices and Intelligent Systems (CODIS), IEEE International Conference on Digital Object Identifier*: Page(s): 318 – 321.

R.J.A. Little and D.B. Rubin.(1987) Statistical analysis with missing data. *Wiley*, New York. pp. 40-43.

D.B. Rubin. (1976), Inference and missing data. *Biometrika*, 63:581–592.

Zadeh L A. (1965) Fuzzy sets. Information and Control, 1965, 8: 338-353

Pawlak Z. (1982) Rough sets. International *Journal of Computer and Information Sciences*, pg: 341-356.

Pawlak Z, Skowron A. (2007), Rudiments of rough sets. *Information Sciences, Elsevier*, 2007, 177 (1) pg: 3-27.

Pawlak Z, Skowron A. (2007), Rough sets: some extensions. *Information Sciences, Elsevier*, 2007, 177 (1): 28-40.

Pawlak Z, Skowron A. (2007), Rough sets and Boolean reasoning. *Information Sciences, Elsevier*, 2007, 177 (1): 41-73.