ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Improvisation on quality of data by handling missing values for mixed attribute data sets

Pravithra CNS[1], Hema MS[2], Sumathi VP[3]

PG Student, Dept of computer science and Engineering, Kumaraguru College Of Technology, Coimbatore, India [1]

Associate Professor, Dept of computer science and Engineering, Kumaraguru College Of Technology, Coimbatore, India [2]

Assistant Professor, Dept of computer science and Engineering, Kumaraguru College Of Technology, Coimbatore, India [3]

[1]pravithrasiva@gmail.com
[2]hema.ms.cse@kct.ac.in
[3]sumathi.vp.cse@kct.ac.in

Abstract: Missing values usually create a noisy environment in all engineering applications and becomes an unavoidable problem in data management and analysis. Various techniques have been developed with great successes on dealing with the missing values in data sets with homogeneous and heterogeneous attributes. For homogeneous data, Sequential method is compared with the special case of the closest fit principle: replacing missing attribute values by most common value from the concept or by filling globally or by mean. This paper studies a new setting of missing data imputation, i.e., imputing missing data in data sets with heterogeneous attributes on local and global class combined with kernel functions, referred to as imputing mixed-attribute data sets. The proposed method is evaluated with the extensive experiments compared with few data sets collected from the UCI repository, and the result demonstrates that the proposed approach is better than the existing imputation methods in terms of classification accuracy.

Keywords: Missing attribute values, sequential and parallel, closes fit, most common value, mean, kernel, data set.

I. INTRODUCTION

Recently on data mining, (i.e.)., discovering knowledge from the raw data, needs receiving a lot of attention. In this paper, the main focus is on missing attribute values which brings a special kind of imperfection and inconsistency. Rules are induced from the given input data sets based on algorithm of sequential methods and kernel functions. Often in real-life data some of the attribute values are *missing* (or *unknown*). Many approaches to handle missing attribute values are studied in [1, 2, and 3].

In this paper the main idea discussed on the closest fit idea. The closest fit algorithm for the missing attribute values is based on replacing the missing attribute value by the existing values of the same attribute were as in another case resembles as much as possible the case with the missing attribute values. In searching of the closest fit case, compare two vectors of attribute values of the given case with the missing attribute values and with searched case and closest fitting cases within the same concept, i.e., among all the cases. The former algorithm as defined is called *concept closest fit*; the latter defined is called *global closest fit*.

[7] Let's assume that input data for data mining are presented in a form of *decision table* (or *data set*) in which *cases* (*records*) are described by the *attributes* (independent variables) and *decision* (dependent variable). The set of all cases with same decision value is called a *concept*. Few theoretical properties of datasets with missing attribute values were studied in [4], [5], and [6]. In general, the methods to handle missing attribute values belong either to *sequential methods* (called *pre-processing methods*) or *parallel methods.*

Imputation of the mixed-attribute data sets can be taken as a new problem in missing data imputation by analyzing discrete and continuous attributes. The challenging issues include such as how measuring on the relationship between instances in a mixed-attribute data set, using the observed data in the data set. To address this issue, the research proposes the nonparametric iterative imputation method based on a mixture kernel for estimating missing values in mixed-attribute data sets. A mixture of kernel functions (linear combination of two single kernel functions, called mixture kernel) is designed such that the mixture kernel is used to replace the single kernel function. The proposed algorithm is experimentally evaluated in terms of classification accuracy with different data sets, such as the nonparametric imputation method with a

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

single kernel and mixed kernel locally and globally. These experiments were conducted on UCI data sets at different missing ratios.

II. RELATED WORK

2.1 SEQUENTIAL METHODS:

Sequential methods actually include techniques based on deleting cases of missing attribute values, replacing missing attribute value by most common value of that attribute, replacing the missing attribute value by mean for numerical attribute values, assigning to a missing attribute value by the corresponding value taken from the closest fit case, or replacing the missing attribute value by a new value, computed from new data set, considering the original attribute as a decision.

In sequential methods the handling of missing attribute values original in-complete data sets, with the missing attribute values, are converted into the whole complete data sets were as the main process, (e.g., rule induction, SVM) is concluded with accuracy [7,8,9,10].

2.1.1 DELETING ALL CASES WITH MISSING ATTRIBUTE VALUES [22]

This method is totally based on ignoring all cases with missing attribute values. It is otherwise called as *list wise deletion* or *case wise deletion.* All the cases with missing attribute values are deleted from the whole data set. Obviously, a lot more information is missing because of deletion. However, there are some reasons [12], [13] to consider it a good method.

2.1.2 MOST COMMON VALUE OF AN ATTRIBUTE [7]

This method is one of the simplest methods to handle missing attribute values, such values are replaced by most common value of the attribute or considered with concept. In other words, a missing attribute value is replaced by most probable known attribute value. Previously, this method of handling missing attribute values is enhanced and implemented, e.g., in CN2 [11].

Let's say that attribute a with missing attribute value for the case x from concept C and value of a for x is missing. Hence this missing attribute value exchanged by known attribute value for which the conditional probability P (*known value of a for case x* $|C$) is the largest. This method already been implemented, e.g., in ASSISTANT [14].

2.1.3 REPLACING MISSING ATTRIBUTE VALUES BY ATTRIBUTE MEAN [7]

This method used for data sets with numerical attributes. In this method, every missing attribute value for the numerical attribute replaced by the arithmetic mean of all known attribute values attribute or considered with concept. The mean of all the known attribute values are replaced in place of missing value instances.

2.1.4 GLOBAL CLOSEST FIT [7]

The global closest fit method [15] based on replacing the missing attribute value by all the known values in another case that resembles as much as possible the case with all the missing attribute values. In searching of the closest fit case, compare two vectors of attribute values, one vector corresponds to the case with the missing attribute value, other vector is a candidate for the closest fit. This kind of search conducted for all the cases, hence the name global closest fit. For each case the distance is computed, the case for which the distance is smallest considered as the closest fitting case used to fill all the missing attribute value.

Let x and y be two cases. The distance between cases x and y is computed as follows,

$$distance(x,y) = \sum_{i=1}^{n} distance(x_i, y_i),$$

where

$$distance(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i, \\ 1 & \text{if } x \text{ and } y \text{ are symbolic and } x_i \neq y_i, \\ & \text{or } x_i = ? \text{ or } y_i = ?, \\ \frac{|x_i - y_i|}{r} & \text{if } x_i \text{ and } y_i \text{ are numbers and } x_i \neq y_i, \end{cases}$$

Where 'r' is difference between highest and lowest of known values of the numerical attribute with missing value. If there is a tie for two cases between with the same distance, a kind of heuristics is necessary.

2.1.5 CONCEPT CLOSEST FIT [7]

This method is much similar to the global closest fit method. The difference is that the original data set, containing missing attribute values, is first split the original data set into the smaller data sets, each smaller data set corresponds to the concept from the original data set. More precisely, every smaller dataset constructed from one of the original concept data set, by restricting the cases to the concept.

2.2 PROPOSED SYSTEM:

In the proposed system the nonparametric iterative imputation algorithm is extended from a single kernel to a mixture of kernels. As discussed in the previous paragraphs about kernel functions, a mixture kernel function is proposed by combining discrete kernel function with a continuous one presented in [15]. Furthermore, a new estimator is constructed based on the mixture kernel by taking local class and global

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

class [16]. The nonparametric iterative imputation algorithm is designed and simply analyzed.

2.3 CLASSIFICATION OF SVM:

Support Vector Machine [7, 8] is a learning machine used as a tool for the data classification, function approximation, etc, due to its main generalization ability, and has found success in many applications. The SVM has been studied extensively for classification, regression and density estimation. Feature of Support Vector Machine is that it minimizes the upper bound of generalization area error through maximizing the margin between the separating of hyper plane and dataset. The SVM classifier actually shows a great performance since it maps the features to a higher dimensional space.

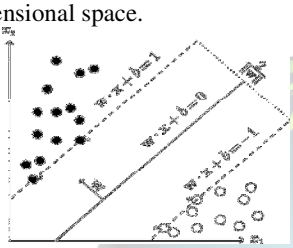


Fig 1: SVM separated Hyper plane

Support vector machine actually constructs a hyper plane or the set of hyper planes in a high or infinite dimensional space, which is mainly used for classification, regression, or other tasks. The hyper plane called functional margin is created, as defined in general the larger the margin the lower the generalization error of the SVM classifier.

[8] To overcome the finite dimensional space, a proposed function introduced that the original finite-dimensional space be mapped into the higher-dimensional space, making the separation easier in space by defining them in the terms of a kernel function [16].

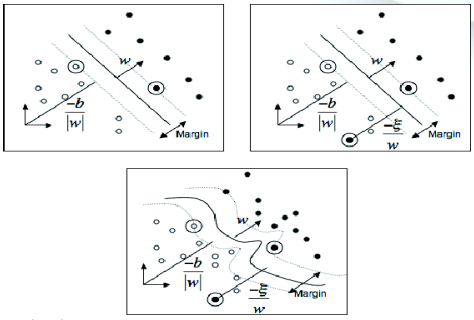


Fig 2: Linear, Non-Linear and Kernel SVM

2.3.1 Polynomial Kernel:

The polynomial kernel is a kernel function that is commonly used with the support vector machines (SVMs) and other kernelized models that represent the similarity of vectors in a feature space over polynomials of the original variables[16].

For degree-*d* polynomials, the polynomial kernel is defined as,

$$K(x, y) = (x^{\mathsf{T}} y + c)^d$$

where *x* and *y* are vectors in the *input space*, i.e. vectors of features computed from training or test samples and $c \geq 0$ is a free parameter. When $c = 0$, the kernel is called homogeneous.

2.3.2 Radial Basis Kernel:

The (Gaussian) or radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms[16]. The RBF kernel on two samples x and x', represented as feature vectors in some *input space*, is defined as

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

$||\mathbf{x} - \mathbf{x}'||^2$ may be recognized as the squared Euclidean distance between the two feature vectors. σ is a free parameter.

III. DESCRIPTION OF DATA SET AND EXPERIMENTAL RESULTS

This datasets are retrieved from UCI Machine Learning. First, a numerical data set of *Breast cancer* data set is taken and analysed with SVM classification. Then the mixed attribute data sets *credit analysis and adult* are taken and analysed with SVM classification. This experiment is implemented using R-Language using packages in R-studio.

3.1 FLOW DIAGRAM:

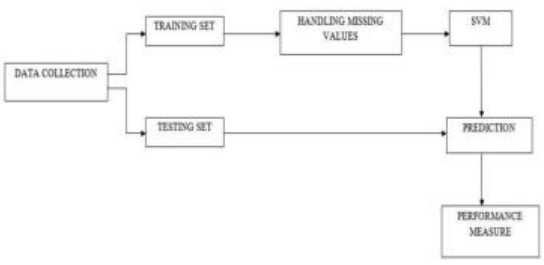


Fig 3: Flow diagram for over existing system

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

3.2 EXPERIMENTAL RESULTS

Initially the results of all the data sets are observed with the R-Studio(R-Language) by using packages like caret, gdata, e1071, Rough sets, etc… The attributes are described below,

S.No	Breast cancer	Credit Analysis	Adult
1.	Sample code number	b, a	age
2.	Clump Thickness	continuous	workclass
3.	Uniformity of Cell Size	continuous	fnlwgt
4.	Uniformity of Cell Shape	u, y, l, t	education
5.	Marginal Adhesion	g, p, gg	education
6.	Single Epithelial Cell Size	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff	marital-status
7.	Bare Nuclei	v, h, bb, j, n, z, dd, ff, o	occupation
8.	Bland Chromatin	continuous	relationship
9.	Normal Nucleoli	t, f	race
10.	Mitoses	t, f	sex
11.		continuous	capital-gain
12.		t, f	capital-loss
13.		g, p, s	hours-per-week
14.		continuous	native-country
15.		continuous	
Class	2 for benign, 4 for malignant	+,-	>50K, <=50K

Table 1 Data set attribute information.

The output from attributes is to predict the class. This is known as the supervised learning.

3.2.1 Pre-processing: SVM classification:

In this section, [9] the pre-processing step is performed to deal with the accuracy of before filling missing values to improve the classification results. Data pre-processing is done to eliminate the incomplete, noisy and inconsistent data. Data must be pre-processed in order to perform any data mining functionality.

SVM Classification

Classification	Breast cancer	Credit Analysis	Adult
Prior Filling	0.9867	0.8118	0.8214

Table 2: Accuracy of Datasets before filling missing values (NA) using SVM

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Kernel Function	0.8118	0.8475	0.7745	0.7892	0.8676	0.9381

Adult Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Kernel Function	0.8214	0.8412	0.7981	0.7991	0.8182	0.8912

Table 3: Accuracy of Datasets of filling missing values (NA) using SVM Kernel function

3.2.2 Global closest fit:

The global closes fit method is based on replacing a missing attribute value by the known value in another case that resembles as much as possible the case with the missing attribute value. By comparing two vectors and measuring its distance.

Breast Cancer Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Global Closest Fit	0.9867	0.9886	0.9826	0.9922	0.9464	0.9783

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Global Closest Fit	0.8118	0.8171	0.7745	0.7747	0.8676	0.8767

Adult Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Global Closest Fit	0.8214	0.8221	0.7981	0.7989	0.8182	0.8213

Table 4: Accuracy of Datasets after filling missing values by global closest fit using SVM

3.2.3 Concept closest fit:

This method is similar to the global closest fit method. The difference is that the original data set, containing missing attribute values split into smaller with respect to concept of cases. For each smaller data set global closest fit is applied and integrated into one.

Breast Cancer Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Concept Closest Fit	0.9867	0.9875	0.9826	0.9916	0.9464	0.9756

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Concept Closest Fit	0.8118	0.8162	0.7745	0.7746	0.8676	0.8678

Adult Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Concept Closest Fit	0.8214	0.8223	0.7981	0.799	0.8182	0.8216

Table 5*:* Accuracy of Datasets after filling missing values by concept closest fit using SVM

3.2.4 Most Common Value:

This method of handling missing attribute values are replaced by most common value of the attribute (i.e.) most probable known attribute value.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Breast Cancer Data Set

Methods	Accuracy		Sensitivity		Specific
	Before Filling	After Filling	Before Filling	After Filling	Before Filling
Most Common Value	0.9867	1.0	0.9826	1.0	0.9464

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specific
	Before Filling	After Filling	Before Filling	After Filling	Before Filling
Most Common Value	0.8118	0.8182	0.7745	0.7767	0.8676

Adult Data Set

Methods	Accuracy		Sensitivity		Specific
	Before Filling	After Filling	Before Filling	After Filling	Before Filling
Most Common Value	0.8214	0.8249	0.7981	0.81	0.8182

Table 6: Accuracy of Datasets after filling missing values by most common value using SVM

3.2.5 Mean:

This method replaces every missing attribute by the arithmetic mean of all known attribute values.

Breast Cancer Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Mean	0.9867	1.0	0.9826	1.0	0.9464	1.0

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Mean	0.8118	0.8207	0.7745	0.7767	0.8676	0.8765

Adult Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Mean	0.8214	0.8279	0.7981	0.8134	0.8182	0.8275

Table 7: Accuracy of Datasets after filling missing values by mean value using SVM

3.2.6 Median:

This method replaces every missing attribute by median of considering known attribute values.

Breast Cancer Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Median	0.9867	1.0	0.9826	0.9	0.9464	0.99

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Median	0.8118	0.8245	0.7745	0.7815	0.8676	0.8712

Adult Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Median	0.8214	0.8256	0.7981	0.7999	0.8182	0.8196

Table 8*:* Accuracy of Datasets after filling missing values by median value using SVM

3.2.7 Deletion Case:

It is used for handling missing values by just deleting all the missing value instances (i.e., NA).

Breast Cancer Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Deletion case	0.9867	0.9868	0.9826	0.9825	0.9464	0.9465

Credit Analysis Data Set

Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Deletion case	0.8118	0.8105	0.7745	0.7748	0.8676	0.8677

Adult Data Set

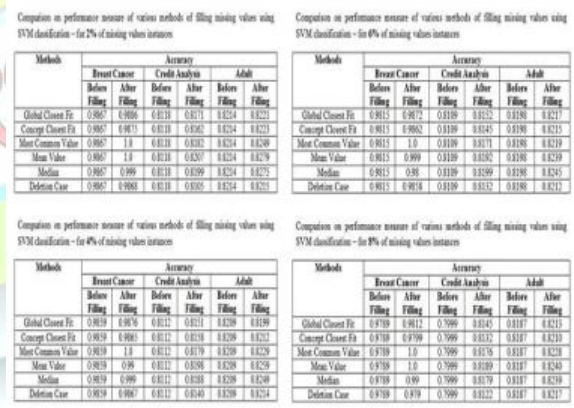
Methods	Accuracy		Sensitivity		Specificity	
	Before Filling	After Filling	Before Filling	After Filling	Before Filling	After Filling
Deletion case	0.8214	0.8215	0.7981	0.7982	0.8182	0.8191

Table 9: Filled missing values using deletion (deleted *NA* instances) and its accuracy using SVM

IV. PERFORMANCE MEASURE

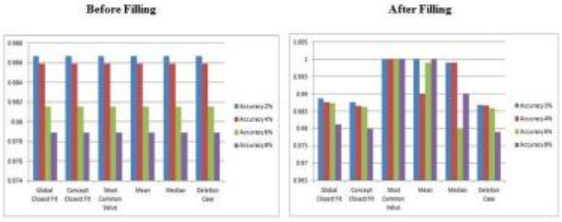
These performance metrics are formally defined as Recall, Precision, F-Measure, and Accuracy.

Table 10: *PERFORMANCE MEASURE*



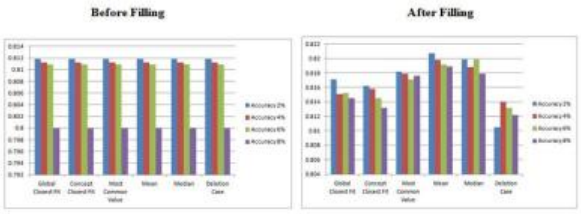
4.1 Graphical representation:

Breast Cancer Data Set



ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Credit Analysis Data Set



Adult Data Set

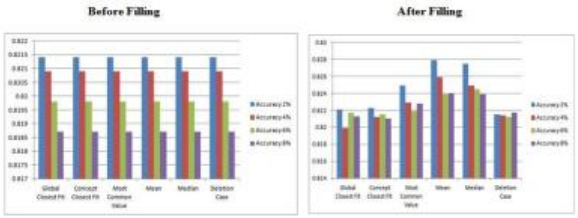


Fig 4*:* Graphical representation for performance measure

V. CONCLUSION

Methods based on multiple ways of imputation on handling missing values coupled with support vector machine classifier is found to be the most suited technique for the imputation of missing values handled to a significant enhancement of different procedures and from the above performance measure predicts the better results for mean and most common value. Therefore Classification accuracy of SVM with kernel function is very much predictable for the missing values imputed by the sequential methods for the mixed attribute data set.

VI. REFERENCES

[1] Grzymala-Busse, J. W.: On the unknown attribute values in learning from examples. Proc. of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina, October 16–19, 1991, 368–377, Lecture Notes in Artificial Intelligence, vol. 542, 1991, Springer-Verlag.

[2] Grzymala-Busse, J.W. and Goodwin, L.K.: Predicting preterm birth risk using machine learning from data with missing values. Bull. of Internat. Rough Set Society 1 (1997) 17–21.

[3] Grzymala-Busse, J.W. and Wang A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. Proc. of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), Research Triangle Park, NC, March 2–5, 1997, 69–72.

[4] Imielinski T. and Lipski W. Jr. Incomplete information in relational databases, Journal of the ACM 31 (1984) 761{791.

[5] Lipski W. Jr. On semantic issues connected with incomplete information databases. ACM Transactions on Database Systems 4 (1979), 262{296.

[6] Lipski W. Jr. On databases with incomplete information. Journal of the ACM 28 (1981) 41{70.

[7] Jerzy W. Grzymala-Busse and Witold J. Grzymala-Busse. Handling Missing Attribute values.

[8] Kristin P. Bennett and Colin Campbell. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explorations ACM SIGKDD, December 2000. Volume 2, Issue 2, page1-* 13.

[9] Qiong Li, Yuchen Fu*, Xiaoke Zhou, Yunlong Xu. The Investigation and Application of SVC and SVR in Handling Missing Values. The 1st International Conference on Information Science and Engineering (ICISE2009) page 1002-1005.

[10] Grzymala-Busse, J.W. and Hu, M. A comparison of several approaches to missing attribute values in data mining. Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing RSCTC'2000, Ban_, Canada, October 16{19, 2000, 340{347.

[11] Clark P. and Niblett T. The CN2 induction algorithm. Machine Learning 3 (1989) 261{283.

[12] Allison P.D. Missing Data. Sage Publications, 2002.
[13] Little R.J.A. and Rubin D.B. Statistical Analysis with Missing Data, Second Edition, J. Wiley & Sons, Inc., 2002.
[14] Kononenko I., Bratko I., and Roskar E. Experiments in automatic learning of medical diagnostic rules. Technical Report, Jozef Stefan Institute, Ljubljana, Yugoslavia, 1984.
[15] Y.S. Qin et al., "POP Algorithm: Kernel-Based Imputation to Treat Missing Values in Knowledge Discovery from Databases," Expert Systems with Applications, vol. 36, pp. 2794-2804, 2009.

[16] Xiaofeng Zhu, Shichao Zhang, Zhi Jin, Zili Zhang, and Zhuoming Xu "Missing Value Estimation for Mixed-Attribute Data Sets" IEEE Transactions on Knowledge and data Engineering, Vol. 23, No. 1, January 2011.