Imputation in Mixed Attribute Datasets using Higher Order Kernel Functions

Aasha.M
Assistant professor
Avinashilingam University

Abstract - The real world data today present have missing values that occur due to a variety of reasons. The methods used for data analyzing such as classification, clustering and dimension reduction procedures require complete data. The missing data present in these must be either removed or preferably estimated . Missing data imputation is a key issue in learning from incomplete data. Many techniques had been developed on dealing with missing values in data sets with homogeneous attributes (their independent attributes are all either continuous or discrete). This paper proposes a higher order spherical kernel based iterative estimator to impute mixed-attribute data sets. The estimators are evaluated with extensive experiments, and the result demonstrates that the proposed approach is better than these existing imputation methods in terms of co relation coefficient and root mean square error (RMSE) at different missing ratios.

Keywords: Imputation, Data Mining, Kernel methods, Mixed Attribute Datasets

I. INTRODUCTION

Many of real world data sets have missing values (MVs) that may occur for a variety of reasons. Methods such as dimension reduction procedures, classification, and clustering require complete data. They must be either removed or, preferably, estimated before such procedures can be employed. To accurately impute missing values , many algorithms have been developed . Missing data imputation[1] aims at providing estimations for missing values by reasoning from observed data. Since missing values can result in bias that impacts on the quality of learned patterns or/and the performance of classifications, imputation of missing values is a major issue in learning from incomplete data. Many methods have been developed to deal with missing values in data sets with homogeneous attributes (their independent attributes are all either continuous or discrete). These techniques cannot be applied to many industrial data sets, and to many real data sets, such as equipment maintenance and gene databases, because these data sets are often with both continuous and discrete independent attributes[18]. These heterogeneous data sets are referred to as mixed-attribute data sets and their independent attributes are called as mixed independent attributes. To meet the above practical requirement, this paper deals with a new method of missing data imputation, which deals with filling of missing data in mixed-attribute data sets using higher order kernel functions.

Imputation of missing values in mixed-attribute data sets can be taken as a recent problem because only few estimators have been designed for this purpose. Major issues in this are, to measure the relationship between instances (transactions) in a mixed-attribute data set, and also to construct mixed estimators using the observed data in the data set. As a solution for this , a nonparametric iterative imputation method based on a mixture kernel is used. In this method, a kernel estimator to find the probability density for independent attributes in a mixed-attribute data set is developed. Secondly, a mixture of two kernel is developed , in which the mixture kernel replaces the single kernel function in the common used estimators. These estimators are called as mixture kernel estimators. Finally, two consistent kernel estimators are constructed for discrete and continuous missing target values, for heterogeneous data sets. Also, a mixture-kernel-based iterative estimator is advocated to use all the available information in complete , as well as observed information in incomplete instances. The experiments were conducted on UCI data sets at different missing ratios. Imputation is the process of finding a feasible or plausible value for a missing value. After imputing all the missing values , the dataset can be analyzed using standard techniques for complete data.. Imputation is not the only method available for handling missing data, but it gives better results when compared to list wise deletion (in which all subjects with any missing values are omitted from the analysis) and may be competitive with a maximum likelihood approach at many times.

II. RELATED WORK

G. Batista and M. Monard [2] analysed four missing data treatment methods for supervised learning. These methods are the 10- NNI method using a k-nearest neighbour algorithm for missing data imputation, the mean or mode imputation, and the internal algorithms used by C4.5 and CN2 to treat missing data. A non-parametric EM- style algorithm for imputing missing values is given by R. Caruana [3]. An iterative nonparametric algorithm for imputing missing values using k-nearest neighbour or kernel regression is proposed instead of parametric models used with EM.

The major step noted is , E and M steps converge into a single step because the data being filled in is the model- updating the filled-in values updates the model at the same time. The major advantages are that (1) it is more accurate for medium size datasets and (2) it is less prone to errors. U.Dick et al [4] gave a learning method with infinite imputation from incomplete data . A generic joint optimization problem in which the distribution governing the missing values is a free parameter was derived. The instantiations of the general learning method consistently outperform single imputations. A mixture models for learning from incomplete data was given by Z. Ghahrami and M. Jordan[5] . their findings gave way to use the present neural network approaches to missing data within a statistical framework and to use a set of algorithms that can take care of clustering, classification and function approximation from incomplete data efficiently.

Imputation of missing network data and some simple procedures were studied by M. Huisman [6]. The various methods are Unconditional mean imputation, Imputation by Reconstruction, Imputation using preferential attachment and Hot deck imputation . From the experiments, reconstruction is found to be the best single imputation method. G. John et al [7] addressed the problem of finding a subset of features that allows a supervised induction algorithm to induce high accuracy concepts. Definitions for irrelevance and for two degrees of relevance were advocated. The understanding behaviour of previous subset selection algorithms was improved by this.. A method for learning Bayesian networks from incomplete datam was given by R. Marco [8] .

The effectiveness of different approaches used for the classification of attributes with unknown values was compared by J.R. Quinlan [9]. The methods used were unknown values when partitioning, unknown values when classifying and unknown values in selecting tests.. J.Racine and Q.Li [10] proposed nonparametric estimation of regression functions with both categorical and continuous data. For bandwidth selection , a data driven method wass also proposed. The new estimator outperforms the conventional nonparametric estimator which has been used to handle the presence of categorical variables. V.C.Raykar and R. Duraiswami [11] gave efficient algorithm for fast optimal bandwidth selection for kernel density estimation.

G.F. Smits and E.M. Jordaan [12] advocated an improved SVM regression using mixtures of kernels. In order to get rid of the disadvantages of single kernels , mixtures of kernels can be used. The performance is also evaluated with artificial as well as an industrial data set. S.C. Zhang et al [13] studied the issue of missing attribute values in training and test data sets. They studied missing data in cost-sensitive learning in which both misclassification costs and test costs are considered.

S.C. Zhang [14] addressed parimputation ie, partially imputation. In this method the missing data is imputed only whenthere is complete data in a small neighborhood of the missing data and, other missing data are not considere in applications such as data mining and machine learning. Missing value methods are categorized as Missing Completely at Random (MCAR),Missing at Random (MAR) and Nonignorable. Practically it is very difficult to meet the nonignorable assumption. MAR is an assumption that is used often , but not always applicable. W. Zhang [15] advocated a technique called association based multiple imputation in multivariate datasets . An outline of all the of association rules are used in multiple imputations in multivariate datasets with missing categorical and numeric data.

III. IMPUTATION TECHNIQUES

This section gives a detailed view about the imputation techniques used in the work namely K-Nearest neighbor, Frequency estimator technique, Imputation using RBF kernel, and Imputation using Poly kernel.

3.1 K –Nearest neighbor imputation method

Missing data imputation is an important step in the process of machine learning and data mining when certain values are missed. Among various imputation techniques, kNN imputation algorithm can be used, as it is a model free and efficient compared with other method. The value of k must be properly taken in this imputation. This subsection deals with the K-Nearest neighbor imputation methods which is explained as follows.

1. Determine parameter K = number of nearest neighbors
2. Calculate the distance between the query-instance and all the training samples
3. Sort the distance and determine nearest neighbors based on the K-th minimum distance
4. Gather the values of Y of the nearest neighbors
5. Use average of nearest neighbors as the prediction value of the query instance

3.2 Frequency Estimation Method

The probability density function of each instance in the row is calculated and the value which is having the highest probability density function is replaced in place of the missing value. This imputation works well in the case of discrete datasets.

3.3 Kernel Functions

Kernel methods are a class of algorithms for pattern analysis or recognition, whose major element used is the support vector machine (SVM).. The level of non-linearity is determined by the kernel function. There are two main types of kernels, namely Local and Global kernels. In local kernels only the data that are close to each other have an influence on the kernel values. whereas, a global kernel allows data points that are not in the proximity from each other to have an influence on the kernel values . In this work polynomial and RBF kernels are chosen The Polynomial and RBF kernels are used again for the analysis of the interpolation and extrapolation abilities. The reason for chosing these two types of kernels is that they can be used as representatives of a broader class of local and global kernels and moreover these kernels have computational advantages over other kernels, since it is easier and faster to compute the kernel.

3.4 Imputation using RBF kernel

An example of a typical local kernel is the Radial Basis Function Kernel (RBF), which is defined in equ(4.1) as follows,

$K(x,x_i)=\exp(-|x-x_i|^2/2\sigma^2$----------------------(3.1)

where the kernel parameter is the width, σ, of the radial basis function. A local kernel only has an effect on the data points in the neighbourhood of the test point. In Figure 3.1 the local effect of the RBF Kernel is shown for a chosen test input, for different values of the width σ[3] .

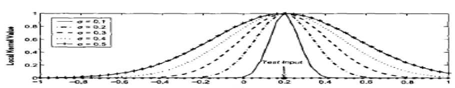

Fig3.1 Local effect of RBF kernel

The missing values are imputed using RBF kernel by choosing a value kernel width σ for which the root mean square error is minimum and correlation coefficient is maximum.

3.5 Imputation using Poly kernel

The polynomial kernel is a typical example of global kernel which is defined in equ (3.2) as follows

$K(x,x_i)=[(x-x_i)+1]^q$---------------------(3.2)

where the kernel parameter q is the degree of polynomial to be used. A global kernel allows data points that are far away from each other to have an influence on the kernel values as well.

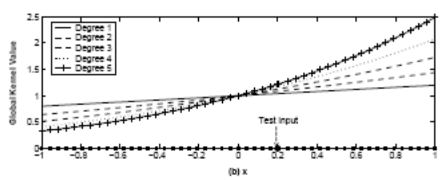

Fig 3.2 Global kernel (Polynomial).

In Figure 3.2 the global effect of the Polynomial kernel of various degrees is shown.

All data points in the input domain have non-zero kernel values for every degree of polynomial. The test data point has a global effect on the other data points. The missing values are imputed using a degree of polynomial 2 to get minimum value of root mean square error and maximum value of correlation coefficient.

IV IMPUTATION USING MIXTURE OF KERNELS

4.1 INTRODUCTION

This section deals with the possibility of using mixture of kernel functions for imputation namely mixing of RBF kernel and Poly kernel and also gives an idea about the mixing of higher order kernel functions like spherical kernel with an RBF kernel and spherical kernel with a Poly kernel for imputing missing data. The mixture of kernel functions when used exhibit better interpolation and extrapolation abilities.

4.2 INTERPOLATION AND EXTRAPOLATION

The quality of a model is not only determined by its ability to learn from the data but also its ability to predict unseen data. These two characteristics are often called learning capacity and generalization ability. Models are seldom good in both of the two characteristics. A typical example is the interpolation and extrapolation abilities of a model. these characteristic are largely determined by the choice of kernel and kernel parameters. Using a specific type of kernel has its advantages and disadvantages. The Polynomial and RBF kernels are used again for the analysis of the interpolation and extrapolation abilities. The reason for analyzing these two types of kernels is that , they can be used as examples of a of local and global kernels respectively and, these kernels have computational advantages over other kernels, since it is easier and faster to compute the kernel values.

It is observed that that for lower degrees of polynomial kernels the extrapolation ability are good, whereas for good interpolation ability higher degree polynomials are required. No single choice of kernel parameter, the degree of the polynomial, results in an estimator for imputation that will provide both good interpolation and extrapolation properties. Similarly, in RBF kernel if large values of σ are used the interpolation ability of RBF Kernels decreases. Therefore, no single value of the kernel parameter, σ, will provide an estimator with both good interpolation and extrapolation properties.

4.3 MIXTURE OF KERNELS

From the previous section, it is observed that a polynomial kernel (a global kernel) shows better extrapolation abilities at lower orders of the degrees, but requires higher orders of degrees for good interpolation. But , the RBF kernel (a local kernel) has good interpolation abilities, but fails to provide good extrapolation. As a result, a mixture of kernel functions namely, RBF kernel and Polynomial kernel can be used for constructing an hybrid estimator for imputing the missing values.
There are several ways of mixing kernels, it is important that the resulting kernel must be an admissible kernel. One way to guarantee that the mixed kernel is admissible, is to use a convex combination of the two kernels Kpoly and Krbf, for example, the mixing of the two kernels is done in the following way as given in equ (5.1). It has been proved experimentally that only a “pinch” of a local kernel, (i.e., 1- ρ = 0.01), needs to be added to the global kernel in order to obtain a combination of good interpolation and extrapolation abilities.

$K_{mix} = \rho K_{poly} + (1- \rho K_{rbf})$------------------(4.1)

where the optimal mixing coefficient ρ has to be determined. In the above equation the value of ρ = constant scalar.

In the techniques done, the values of, ρ, q, and σ are combined with the coefficient λ in such a way so that the value of the Approximate Mean Integrated Square Error(AMISE) is optimized. First, the value of ρ is limited. If the data are in a (0, 1) scaled input space, a pure RBF-kernel with ρ > 0:4 behaves like a lower degree polynomial in the known learning space.

The RBF-kernel is needed when the local behavior is modelled. Using a value of λ that is too small will result in complex models that also models the noise. So, it is appropriate to set the value λ of between 0.15 and 0.3 in this approach. It is also noted that polynomial kernel , which is a global kernel has good extrapolation abilities . This extrapolation abilityand shows sudden increases or decreases when the value of q is too high. In order to avoid this, a lower degree for the polynomial kernel is taken. The value of, d > 2 is rarely used, and q is usually set to 1 or 2. The choice of ρ is depends upon the amount of local behavior which needs to be modeled by the RBF kernel. In the experiments, it is better if to chose the value of ρ between 0.95 and 0.99.

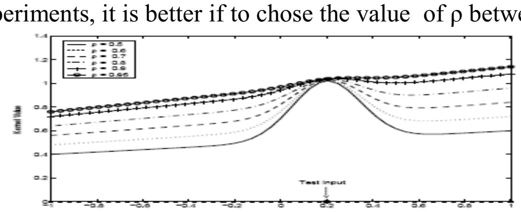

Fig 4.3 Example of a mixed kernel

Figure 4.3, shows the effect of mixing a Polynomial kernel with a RBF Kernel .

Two consistent kernel functions one for discrete and one for continuous are designed and the two functions are combined to design a mixture kernel function for imputation which is defined in equ (4.2) as follows.

$K_{h,\lambda,\, ix} = K(^{x- x}{}_{i/h})\, L(X_{id,}\ x_{id}\, , \lambda)$------------------------(4.2)

where h → 0 and λ → 0 (λ; h is the smoothing parameter for the discrete and continuous kernel functions, respectively), and Kh; λ;i_x is a symmetric probability density function. Using the above function , the missing values are imputed for the datasets Abalone, Housing, Import, Pima and Vowel and the performance is evaluated using root mean square error and correlation coefficient.

4.4 IMPUTATION USING SPHERICAL KERNEL AND RBF KERNEL

In this technique the imputation of missing values is done using a mixture of higher order kernel function namely spherical kernel and RBF kernel and also spherical kernel with a polynomial kernel . The spherical kernel is a higher order kernel which has higher rate of computation and it is defined in equ (4.3)as follows

$$K(x, xi) = 1-3/2\,\|x-xi\|/\sigma + ½\,(\|x-xi\|/\sigma)\text{-----------}(4.3)$$

The mixing of spherical kernel with RBF kernel is done in the following way and the new kernel function is defined in equ (4.4) as follows

$$Kmix = \rho Ksph + (1- \rho\, Krbf)\text{--------------------}(4.4)$$

In this method , ρ is a constant scalar and the best values of λ, σ, q and d should be combined. This can be obtained by experimenting all the combinations of σ, q, λ, and d, so that the complexity is reduced.

4.5 IMPUTATION USING SPHERICAL KERNEL AND POLY KERNEL

As described in the previous section , the imputation is also done using a mixture of spherical kernel[8] and Poly kernel[2] which is defined in equ (4.5) as follows

$$K\ mix = \rho\, K\ sph + (1- \rho\, Kpoly)\text{----------------}(4.5)$$

Using the above kernel function an estimator is created for imputation of missing values and the performance is evaluated for all the techniques mentioned above, with respect to root mean square error.

V EXPERIMENTAL SETUP AND DATASET DESCRIPTION

This section deals with the experimental set up used and also the details about the datasets used in the experiments.

5.1 DATA SET ACQUISITION

Five publicly available data sets are used in this work namely Imports, Housing, Abalone , Pima and Vowel. These data sets are obtained from UCI machine learning repository and these data sets does not contain any missing values. The missing values are created at random in these datasets and are imputed using the techniques used in this work. Finally the performance of these techniques is evaluated using root mean square error and correlation coefficient.

5.2 DATA SET DESCRIPTION

The data sets used in the experiments are mainly obtained from UCI machine learning data repository. Table 5.1 gives details about the datasets used in this work.

Name	**Type**	**#(attr.)**	**#(ins.)**
Imports	C	8	398
Housing	C	14	506
Abalone	D(29)	8	4177
Pima	D(2)	8	768
Vowel	D(11)	10	528

Table 5.1 Datasets used in the experiments

In the above table 5.1, the first column indicates the name of the data sets used, the second column denoting type indicates whether it is C (continuous) and D(discrete) and the third column indicates the number of the attributes and the last column denotes the number of instances in these data sets.

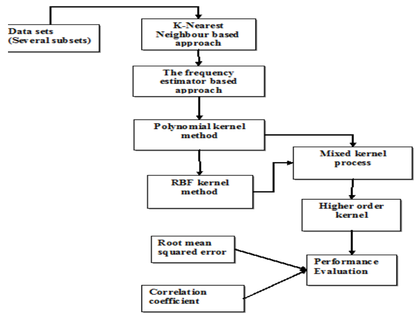

Fig 5.1 Data flow diagram

Fig 5.1 gives an overview of the work. The original data sets are subjected to pre processing techniques and the missing values are imputed using K-NN, Frequency estimator method, RBF kernel and Polynomial kernel and then a mixed kernel (RBF kernel and Poly kernel), and a spherical kernel mixed with poly kernel and a spherical kernel mixed with RBF kernel. Finally the performance of these imputation methods is evaluated using Root Mean Square Error(RMSE) and Correlation Coefficient.

VI RESULTS AND DISCUSSION

6.1 INTRODUCTION

This chapter deals with the performance evaluation and describes about the metrics used for evaluating the performance of the techniques used in this work. The performance is evaluated against root mean square error and correlation coefficient. The root mean square error and correlation coefficient are calculated for all the five data sets namely Abalone, Import, Housing, Vowel and Pima and results are compared.

6.2 PERFOMANCE EVALUATION

The performance is evaluated using Root mean square error (RMSE) and correlation coefficient.

$RMSE= \sqrt{1/m} \sum_{i=1}^{m}(e_i - \sim e_i)^2$ ---------------(6.1)

Equ(6.1) gives the expression for calculating RMSE,where

e_i = the original attribute value;

$\sim e_i$ = the estimated attribute value,

m = the total number of predictions.

When the value of the RMSE is larger, the prediction will be less acurate.

The correlation coefficient is usually defined as follows

$P_{x,y} =corr(x,y)= cov(x,y)/\sigma_x \sigma_y$--------------------(6.2)

Equ(6.2) gives the expression for calculating correlation coefficient , where x denotes the imputed value, y denotes the observed value and cov denotes the covariance and corr denotes the correlation between the two values.

6.3 EXPERIMENTAL RESULTS AND DISCUSSION

The root mean square error and correlation coefficient are calculated for all the techniques discussed and the performance is compared. The RMSE value and correlation coefficient are calculated for all the datasets as discussed above. Table 6.1 gives the results of RMSE and correlation coefficient for Import data set

S.NO	Name of the method used for Imputation	Auto-Import data set	
		RMSE(%)	Correlation coefficient(%)
1	K-NN	0.847193	0.650316
2	Frequency estimator	0.809691	0.719151
3	RBF kernel	0.673918	0.791693
4	Poly kernel	0.730813	0.743009
5	Mixed kernel(RBF and Poly)	0.653771	0.828723
6	Spherical and RBF	0.603390	0.900268
7	Spherical and Poly	0.564306	0.939870

Table 6.1 RMSE and correlation coefficient for Auto-Import data set

The Table 6.2 given below shows the results for Housing data set

S.NO	Name of the method used for Imputation	Housing data set	
		RMSE(%)	Correlation coefficient(%)
1	K-NN	**0.83133**	**0.67677**
2	Frequency estimator	**0.80052**	**0.68361**
3	RBF kernel	**0.67604**	**0.78596**
4	Poly kernel	**0.76276**	**0.73970**
5	Mixed kernel(RBF and Poly)	**0.65959**	**0.84034**
6	Spherical and RBF	**0.54589**	**0.94115**
7	Spherical and Poly	**0.57777**	**0.92622**

Table 6.2 RMSE and correlation coefficient for Housing data set

Similarly the results for all the five datasets can be obtained and it is observed that from the above results that spherical kernel which is a higher order kernel when mixed with Rbf kernel and Poly kernel gives a least value of RMSE for all the data sets. It is also observed that correlation coefficient is maximum when imputation is done using a mixture of spherical kernel with Rbf and spherical kernel with Poly Kernel. Thus it is observed that spherical kernel when mixed with Rbf kernel and spherical kernel when mixed with Poly kernel produces better results. Fig (6.1) And Fig(6.2) gives the values of RMSE and Correlation Coefficient for all the five datasets when missing values are imputed in these five datasets using the five methods as described in the previous sections.

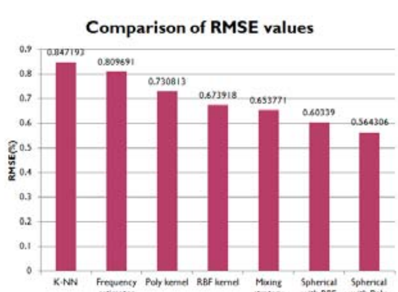

Fig 6.1 Comparison of RMSE values for the datasets

Fig(6.1) shows the different RMSE values for all the five datasets when imputation is done using the techniques as discussed in previous sections. Similarly Fig(6.2) shows the different CC values for all the five datasets , when they are imputed using all the above techniques as discussed previously.

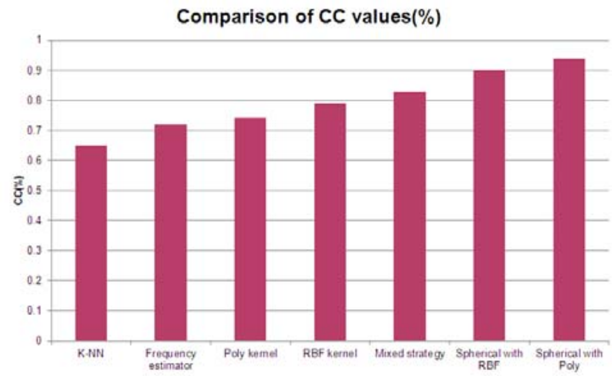

Fig 6.2 Comparison of CC values for the Data Sets

It can be observed from these figures that imputation using mixture of Spherical and Poly kernel and mixture of Spherical and RBF kernel produces good results.

VII CONCLUSION

In this work, a mixture kernel estimator using spherical kernel and Rbf kernel and also using spherical kernel and Poly kernel has been proposed for imputing missing values in a mixed-attribute data set. This mixture kernel-based iterative nonparametric estimator is proposed against the case that data sets have both continuous and discrete independent attributes. The future work can be planned to further explore other kernel functions,in order to obtain better extrapolation and interpolation abilities in learning algorithms. Other higher order kernel functions can be explored in the future to produce better results.

REFERENCES

[1] J.Han and M.Kamber, Data Mining Concepts and Techniques, second ed. Morgan Kaufmann Publishers,2006.

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

[4] U. Dick et al., “Learning from Incomplete Data with Infinite Imputation,” Proc. Int’l Conf. Machine Learning (ICML ’08), pp. 232-239, 2008.

[5] Z. Ghahramani and M. Jordan, “Mixture Models for Learning from Incomplete Data,” Computational Learning Theory and Natural Learning Systems, R. Greiner, T. Petsche, and S.J. Hanson, eds., vol. IV: Making Learning Systems Practical, pp. 67-85, The MIT Press, 1997.

[6] M. Huisman, “Missing Data in Social Network,” Proc. Int’l Sunbel Social Network Conf. (Sunbelt XXVII), 2007.

[7] G. John et al., “Ir-Relevant Features and the Subset Selection Problem,” Proc. 11th Int’l Conf. Machine Learning, W. ohen and H. Hirsch, eds., pp. 121-129, 1994.

[8] R. Marco, “Learning Bayesian Networks from Incomplete Databases,” Technical Report kmi-97-6, Knowledge Media Inst., The Open Univ., 1997.

[9] J.R. Quinlan, “Unknown Attribute values in Induction,” Proc.Sixth Int’l Workshop Machine Learning, pp. 164-168, 1989.

[10] J. Racine and Q. Li, “Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data,”J. Econometrics, vol. 119, no. 1, pp. 99-130, 2004.

[11] V.C. Raykar and R. Duraiswami, “Fast Optimal BandwidthSelection for Kernel Density Estimation,” Proc. SIAM Int’l Conf.Data Mining (SDM ’06), pp. 524-528, 2006.

[12] G.F. Smits and E.M. Jordaan, “Improved SVM Regression Using Mixtures of Kernels,” Proc. 2002 Int’l Joint Conf. Neural Networks,pp. 2785-2790, 2002.

[13] S.C. Zhang et al., “Missing Is Useful: Missing Values in Cost-Sensitive Decision Trees,” IEEE Trans. Knowledge and Data Eng.,vol. 17, no. 12, pp. 1689-1693, Dec. 2005.

[14] S.C. Zhang, “Parimputation: From Imputation and Null-Imputationto Partially Imputation,” IEEE Intelligent Informatics Bull.,vol. 9, no. 1, pp. 32-38, Nov. 2008.

[15] W. Zhang, “Association Based Multiple Imputation in MultivariateData Sets: A Summary,” Proc. Int’l Conf. Data Eng. (ICDE),p. 310, 2000.

[16] A. Dempster, N.M. Laird, and D. Rubin, “Maximum Likelihoodfrom Incomplete Data via the EM Algorithm,” J. Royal Statistical Soc., vol. 39, pp. 1-38, 1977.

[17] D. Chao Feing, Z. Wang and J.Fang Shi, “Research on Missing Value Estimationin Data Mining” Proceedings of the 7th World Congress on Intelligent Control and AutomationJune 25 - 27, 2008, Chongqing, China

[18] X. Zhu, S. Zhang, Z. Jin, Z.Zhang and Z. Xu: “Missing Value Estimation for Mixed Attribute Data Sets”, IEEE Transactions on Knowledge and Data Engineering.,vol. 23, no. 1, Jan 2011

[19] http://crsouza.blogspot.in/2010/03/kernel-functions-for-machine-learning.html