An Estimation of Missing Values by Modified Mixed Kernels

Hemalatha.S, and M.Hemalatha

***Abstract*-----**In statistical practices, difficulties of missing data are universal. Several techniques are used to handle this dilemma of missing data. They include both old approaches, which require only a small amount of mathematical computations and new approaches, which require additional difficult computations that are ever easier for social work researchers to carry out the statistical programming softwares. In the existing system, there is a novel setting of missing data imputation, i.e. imputing in mixed-attribute data sets. This system offers two consistent estimators for discrete and continuously missing target values, correspondingly. After that a mixture-kernel based iterative estimator is offered to impute mixed-attribute data sets. In this method, the local kernel and global kernel are used and linear combination of these mixed kernels is used. Nevertheless, the accuracy of the system is decreased with the large number of data samples. Unquestionably it will degrade the performance of the system. To improve the performance and to increase the accuracy of the system we proposed three approaches. First we introduce the local kernal RBF using KL divergence, secondly we introduce the global kernal polynomial using probability distribution and finally mixed kernels in piece level combination instead of linear combination. From the experimental result we can obtain that the proposed system is much more effective than the existing system. The performance also is shown to have improved in this proposed system.

***Keywords*----**Missing data imputation, Kernel Function Selection, Linear Mixture Kernel Function, RBF kernel, Polynomial kernel and Statistical Imputation for Missing Data.

I. Introduction

IMPUTATION of Missing data [1] is the important objective as it offers to find out the missing values by calculation from experimental data. Since the missing values lead to less accuracy that in turn result in the worthiness of the patterns and/or the efficiency of the classification, estimation of missing data has been an important issue in learning from incomplete data. In the datasets with homogeneous attributes, which are described as independent attributes, numerous numbers of techniques are being built up for the missing values estimation. Despite the fact that these imputation techniques cannot be supported to many real data sets, for instance equipment maintenance databases, gene databases, and industrial data sets [2], these data sets are frequently with both continuous and discrete independent attributes. The heterogeneous data sets are generally named as 'mixed-attribute data sets'. In these heterogeneous data sets have independent attributes which are called 'mixed independent attributes'.

Imputation of mixed attribute data sets is a new issue in the imputation of missing data, since there is no estimator proposed for imputing missing data in mixed attribute data sets. Some of the difficult problems are, for instance, how to measure the correlation between instances in a mixed-attribute data set and how to build the novel estimators using the observed data in the data set. To overcome this difficulty, this system proposes a nonparametric iterative imputation approach based on mixture kernels which are used for missing values imputation in mixed-attribute data sets. To understand the probability density for independent attributes, a kernel estimator is built first. After that, a mixture of kernel functions, which is defined as a linear combination of two single kernel functions, is proposed for the estimator. In this estimator the mixture kernel is used as a replacement for the single kernel function in the traditional kernel estimators. This estimator is called 'mixture kernel estimator'. According to this, two consistent kernel estimators are built for discrete and continuous missing target values, correspondingly.

The mixture kernel based iterative estimators make use of all accessible observed information. The available information contains the observed information in incomplete instances (with missing values), to impute missing values, from observed information which contains complete instances (without missing values). To defeat this issue various applications has been used to perform missing value imputation. It is a specific and difficult issue confronted by machine learning and data mining techniques. Consequently, there are many challenges in the imputation of missing value. The established missing value estimation approaches can be usually classified as regression imputation (RI) and the Nearest Neighbour Imputation (NNI) which is demonstrated in [3]. Following this, by replacing them with some believable values, missing values in a dataset are concluded. The sensible values are normally produced from the dataset using an imputation method.

Of late, imputing performed in mixed-attribute data sets poses to be a significant difficulty in the missing value imputation. In view of the fact that, for estimating the missing data in mixed attribute data sets, the estimator is not considered. The challenges include issues such as how measuring the association between instances in a mixed-attribute data set. There is also a concern as to how to build the hybrid estimators using the observed data in the data set.problem.

Hemalatha.S, and M.Hemalatha, are with Department of Computer Science, Karpagam University, Coimbatore, TN, India. (hemamca_2006@yahoo.co.in1 and csresearchhema@gmail.com2)

II. Related Works

Approaches for Missing value are categorized into three methods. These ideas are offered in the following works such that [5], [6], [7]: i) case deletion, ii) learning without handling of missing values, and iii) missing value imputation. In the first approach, case deletion [8] is a method which is used basically to ignore those cases that come along with the missing values and to complete the learning progression only to utilize the residual instances. For the second technique of the learning without handling of missing data schemes, such as Bayesian Networks method, Artificial Neural Networks method, and some more approaches are demonstrated in [9], [10]. The third approach is entirely dissimilar to the above two techniques. This approach maintains filling in missing values. This producer is processed before the learning procedure starts. Missing data imputation is nothing but an approach which is used to replace the missing values with other practical values, such as the ones presented in [11] and [12]. Though the imputation method is considered as a more favourable approach [13], a new research direction, the Parimputation (Partially Imputation) scheme, has been introduced recently in [14]. It supports the imputation of a missing datum. If some complete instances in a small neighborhood of the missing datum are there, then only this data are imputed, if not that missing data is not imputed. By using the grid search technique, the optimal bandwidth is selected in the mixture kernel estimators, which is a replacement of the data-driven technique used in [15].

Q.H. Wang and R. Rao [11], proposed a empirical likelihood technique which is benefits to the missing value response difficulty.

The important aim of this research is to expand the empirical likelihood technique to the missing response trouble, which was well thought-out by Cheng; in addition, it also aims at making assumptions on the mean of response *Y*. The overall idea is to estimate the missing *Y* –values by the kernel regression imputation and then to build the complete data empirical likelihood for θ, which is obtained from the imputed data set as if they were independent and identically distributed (i.i.d.) observations. However, the imputed data are not i.i.d. Significantly, the empirical log-likelihood ratio in the estimation is asymptotically dispersed as a scaled chi-square value. Nevertheless, the empirical log likelihood ratio cannot be functional straight to construct statistical inference on θ. Therefore, adjustment of the empirical log-likelihood ratio is one main motivation for this research. In other words, the adjusted log likelihood ratio is asymptotically dispersed as a standard chi-square value. It is well-known that Adimari (in (1997) used the empirical likelihood approach, but to build inference under random restriction and acquired an analogous result. By making effective use of the known auxiliary information on *X* and $\widehat{\theta_n}$ the empirical likelihood method, an empirical likelihood-based estimator of θ is proposed, which has a smaller asymptotic variance than , and some truncated versions of it. Besides, an adjusted empirical likelihood ratio with auxiliary information is also attained and then it is applied to make confidence intervals for θ.

V.C. Raykar.et.al,[16] Propose a computationally efficient -*exact* approximation algorithm. This is proposed for univariate Gaussian kernel based density derivative estimation. This algorithm reduces the complexity of computational from $O(MN)$ to $O(N+M)$; in other words the computational complexity is reduced from $O(N^2)$ to $O(N)$. In this work, the scheme is applied to calculate approximately the optimal bandwidth, which is used for kernel density estimation. This paper also presents a process with a speedup obtained for optimal bandwidth estimation. This estimation method supports both simulated data and real data. In this work the Taylor's series expansion is used about a certain point x^*. On the other hand, when the same x^* is used for all the points, generally it would process the high truncation number p for the reason that the Taylor's series offers good estimation only in a small open interval around x^*. Consistently the space is sub-divided into K intervals with the length $2rx$. The N source points are given into K clusters, Sn for $n = 1;. . . . . ;K$ with cn being the center of each cluster. After this process completion, the collective coefficients are estimated for each cluster and then summations of all clusters's total contribution are performed. As the Gaussian decays very fast, acceleration is acquired if all the sources belonging to a cluster are omitted and if the cluster is greater than a certain distance from the target point.

III. Existing Methodology

In this imputation approach, the ith missing value is denoted by MV_i and the imputed value of MVi in tth iteration imputation is regarded as $\widehat{M}V_i^t$. From the above algorithm, all the imputed values are used to impute subsequent missing values. This means that the (t + 1)th (t - 1) iteration imputation is performed according to the imputed results of the tth imputation, up till the filled-in values converge or start on to cycle or satisfy the requirements of the users.

In this first iteration of imputation process, all missing values are imputed, including using the mean for continuous attributes. In the field of machine learning, replacing the missing values with the mean of an attribute is a well-liked imputation technique. Nevertheless, Brown pointed that this method is applicable if and only if the data set is taken from a population with a normal distribution. This is typically impracticable for real applications for the reason that the real distribution of a data set is can't predetermined. Conversely, Rubin established that a single imputation cannot offer valid standard errors and confidence intervals, given that it disregards the uncertainty implicit in the fact that the imputed values are not the actual values. Hence, according to the first imputation, extra iteration-imputations performed are reasonable and essential for improved dealing with the missing values.

Analysing the second iteration of imputation, each of them is carried out on the basis of the former imputed results, with the nonparametric kernel estimator. During the imputation process, when the missing value $\widehat{M}V_i^t$ is imputed

according to Iterative Kernel Estimator for Continuous Target Variable and Discrete Target Variable, all other missing values are regarded as observed values, i.e., MVi = $\widehat{M} V^{t-1}_i$, p ε S_m, p = 1, . . . ,m, p ≠ i. In particular, $\widehat{M}$ V^1_i = mean (S^r in Y) if the target variable Y is a continuous variable, $\widehat{M}$ V^1_i = mode (S^r in Y) if Y is a discrete one in this algorithm. While the filled-in values start converging or begin a cycle, the missing continuous attributes imputation iteration will be finished. For discrete missing values, the imputation algorithm will be terminated if $|CA_t - CA_{t-1}| \geq \varepsilon$, based on the principle of the parameter iterative algorithm EM, where ε is a nonnegative constant specified by users; the classification accuracy for the tth imputation is denoted by CA_t. After completing the first imputation, the time of iteration of the algorithm is t for discrete missing attribute imputation.

Zheng et al. showed that Polynomial kernel (such as the global kernel) has good extrapolation at lower order degrees; however it needs higher order degrees for obtaining a well interpolation. Also RBF kernel (such as local kernel) is well in interpolation, although it does not succeed to offer good extrapolation. They also showed that a mixture of kernels can lead to much better extrapolation and interpolation than using either the local or global kernels. In this study, the proposed imputation technique is founded on a mixture kernel function which is defined as follows:

Linear Mixture Kernel Function: Let Kpoly = (< x, xi > +1)^q, Krbf = exp(-(x – xi)^2 / σ^2), a linear mixture kernel function is defined as follows:

Kmix = ρ Kpoly + (1 - ρ) Krbf

where q is the degree of the polynomial, σ is the width of the radial basis function (RBF), and ρ is the optimal mixed coefficient (0 ≤ ρ ≤ 1). The values of ρ, q, and σ are constant scalars, but have to be determined with experiments.

For imputing missing data in a mixed-attribute data set the consistent kernel regression is proposed. This proposed mixture kernel- based iterative nonparametric estimator supports both the data sets cases, i.e. both continuous and discrete attributes. This system makes use of all accessible observed information. It make use of all available observed information, as well as observed information in incomplete instances which contains missing values for imputing the missing values, whereas existing imputation techniques develop the system which is considered only the observed information in complete instances which is not contains the missing values. The optimal bandwidth is choosen by use of grid search approach.

IV. Proposed Methodology

In the proposed system, three novel approaches are introduced. They are:

1. RBF kernel using KL divergence
2. Polynomial kernel using probability distribution
3. Piece level combination in mixed kernels

In the following section, a detailed description of these three approaches is presented, with a view to pronouncing the novelty in their performance. These three techniques are very well used in the proposed system to improve accuracy. The following chapters clearly explain how these techniques improve the performance of the system.

1. RBF kernel using KL divergence

The (Gaussian) radial basis function kernel (RBF), is a well-liked kernel function used in support vector machine classification. In general, in the machine learning fields it is widely used. The RBF kernel on two samples x and x', represented as feature vectors in some *input space*, is defined as

$$K(x, x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$$

Where, $\|x-x'\|_2^2$ is defined as the squared Euclidean distance between the two feature vectors.

σ is a free parameter.

In the proposed system, in order to improve the performance of the system, the RBF kernel using Kullback-Leibler divergence (KL-divergence) is introduced. In general, RBF kernel is derived from the Euclidean distance measure. Instead of Euclidean distance between the two feature vectors, we are proposed the KL divergence between the two feature vectors. KL divergence is a popular measure for a similarity between two probability distributions. The KL-divergence is defined as,

$$KL\,[p \parallel q] = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

Where, p and q are two feature vectors in some *input space*.

The proposed RBF kernel using KL divergence is defined as follows,

$$K(x, x') = \exp\left(-\frac{\|KL\,[x \| x']\|_2^2}{2\sigma^2}\right)$$

Where, $\|KL\,[x \parallel x']\|_2^2$ is defined as the squared KL divergence between the two feature vectors

An equivalent, but simpler, definition involves a parameter: $\gamma = -\frac{1}{2\sigma^2}$

$$K(x, x') = \exp\left(\gamma \|KL\,[x \parallel x']\|_2^2\right)$$

Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one (when **x** = **x'**), it has a ready interpretation as a similarity measure.

2. Polynomial kernel using probability distribution

The polynomial kernel is also a kernel function which is mainly used in support vector machines (SVMs) and other kernelized models. This kernel function represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models, in machine learning fields.

Spontaneously, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the framework of regression approaches, such combinations are known as interaction features. The polynomial kernel's feature space (implied) is equivalent to that of polynomial regression, however, without the combinatorial blow up in the number of parameters to be learned. While the input features are binary-valued, they match up to logical conjunctions of input features.

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

$K(x, y) = (x^T y + c)^d$

where x and y are vectors in the input space, in other words vectors of features computed from training or test samples, $c \geq 0$ is a constant trading off the influence of higher-order versus lower-order terms in the polynomial. The kernel is such as homogeneous while the c = 0.

To prove that the proposed system ensures an improved accuracy, the probability distribution in the polynomial kernel is introduced. The proposed polynomial kernel function is derived with the help of the probability distribution of feature vectors in the input space. The proposed polynomial kernel function is built by the multiplication of the probability distribution of two feature vectors. Thus, it is defined as,

$K(x, y) = (p(x)\, p(y) + c)^d$

Where, p(x) is probability distribution of x feature vectors p(y) is probability distribution of y feature vectors

3. Piece level combination in mixed kernels

Polynomial kernel (such as the global kernel) has good extrapolation at lower order degrees; however it needs higher order degrees for obtaining an improved interpolation. Also RBF kernel (such as local kernel) is well in interpolation, although it does not succeed to offer good extrapolation. To obtain a stronger extrapolation and interpolation in the existing system mixed kernels are proposed. They also showed that a mixture of kernels can lead to much better extrapolation and interpolation than using either the local or global kernels.

To achieve much better extrapolation and interpolation abilities, piece level combinations of the kernels are used. In other words, to improve the extrapolation and interpolation, piece level combination based mixed kernels are introduced. In this proposed approach, the global and local kernels are mixed in the piece level. It increases the accuracy of the system. The results of Mixing, RBF, and Polynomial kernels show that the effects of the Mixing algorithm are better than the single kernel. In other words, using mixture kernels in nonparametric kernel estimation can offer good learning facility and generalization capability compared to the single kernels (either the RBF or polynomial kernels) estimation.

V. Experimental Results

Dataset description

In the experiment conducted for the proposed study, UCI data sets were taken. There are two types of datasets in the UCI data viz. continuous data and discrete data. From the continuous data, two datasets were taken namely Auto-mpg and Housing. From the discrete data, four datasets namely Abalone, Pima, Vowel and Anneal were taken. In the continuous data, Auto-mpg has the attribute as categorical and real attributes. This dataset consists of 398 instances and 8 attributes. And Housing continuous data has attribute as categorical, integer and real attributes. This dataset contains 506 instances and 14 attributes.

In the discrete data, Abalone has attributes such as categorical, integer and real attributes. It contains 4177 instances and 8 attributes. The attributes of Anneal data are also categorical, integer and real attributes. It consists of 798 instances and 38 attributes. In Pima data, the type of attributes is integer and real. This data contains 768 instances and 8 attributes. The Vowel discrete data has only real attributes and also it contains 640 instances and 12 attributes.

Performance evaluation results

In this section, a comparison of the performances between the existing systems such as KNN (K- Nearest Neighbour), FE (Frequency Estimator), RBF (Radial basis function kernel), Poly (polynomial kernel) and Mixed kernels and the proposed system i.e., Modified mixed kernel. The performance of the proposed system is evaluated in terms of the parameter such as correlation coefficient, RMSE (Root Mean Square Error) and classification accuracy.

CC is defined as the strength and direction of a linear relationship between two variables. In our system, the correlation coefficient between the actual and predicted values of missing attributes is calculated after convergence. The CC value is +1 in the case of a perfect increasing linear relationship and -1 in case of a decreasing linear relationship, and the values in between point out the degree of linear relationship between for actual and predicted values of missing attributes. The CC of 0 means that there is no linear relationship between the attributes.

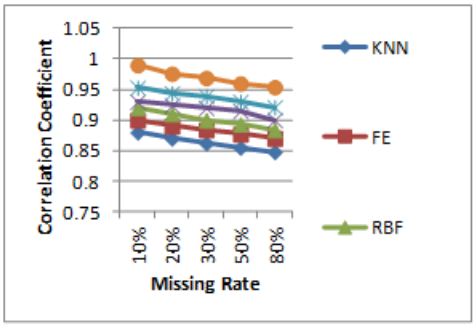

Fig.1 Correlation Coefficient comparison

In the above graph, the performance comparison is shown between the existing system such as KNN, FE, RBF, Poly and Mixed kernels and proposed system i.e., Modified mixed kernel in terms of CC (Correlation Coefficient). In this graph, the x axis will be the missing rate and the y axis will be the CC. For different missing rate we are calculating the CC for six approaches. While the missing rate is increased, CC vale is decreased correspondingly. The proposed system has high CC compared to the other existing system. From this, it can be easily concluded that the proposed system is well effective in CC parameter.

The RMSE (also called the root mean square deviation, RMSD) is a commonly used measure of the difference between original attribute value and estimated attribute value. The RMSE of an estimated attribute value with respect to the original attribute value is defined as the square root of the mean squared error:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m}(e_i - \tilde{e}_i)^2}{m}}$$

where e_i is the original attribute value; $\tilde{e}_i$ is the estimated attribute value, and m is the total number of predictions. The accuracy of the system depends on this RMSE rate value because, if the RMSE rate is larger, then the accuracy of the system decreases.

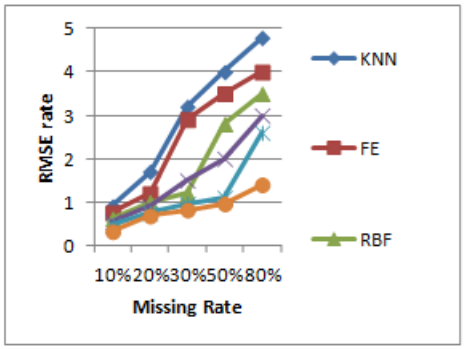

Fig.2 RMSE comparison

The above graph shows the performance comparison between the existing system such as KNN, FE, RBF, Poly and Mixed kernels and proposed system i.e., Modified mixed kernel, in terms of the RMSE rate. In this graph, the x axis will be the missing rate and the y axis will be the RMSE rate. For different missing rate the RMSE rate is calculated for six approaches. While the missing rate is increased, RMSE rate also increases correspondingly. The proposed system has a low RMSE rate compared to the other existing system. From this it can be easily understood that the proposed system is very effective in RMSE rate.

In the section, the performance of the proposed system is evaluated in comparison with the existing system in terms of classification accuracy measurement. The Accuracy can be calculated from the formula given as follows:

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

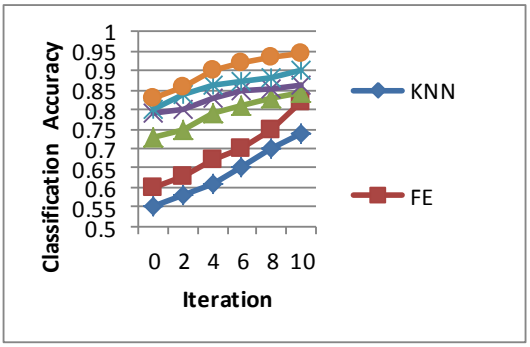

Fig.3. Classification Accuracy comparison

The above graph shows that performance comparison between the existing system such as KNN, FE, RBF, Poly and Mixed kernels and proposed system i.e., Modified mixed kernel in terms of classification accuracy. In this graph the x axis will be iteration and the y axis will be classification accuracy. In different iteration, the classification accuracy is calculated for six approaches. While the iteration is increased, classification accuracy also increased correspondingly. The proposed system has a higher classification accuracy compared to the other existing system. From this it can be easily inferred that the proposed system is very effective in classification accuracy.

VI. Conclusion

Missing data is an everyday dilemma in economics, since the variables missing from a data set or values lead to missing observations. In the existing system, consistent kernel regression was proposed for imputing missing data in a mixed-attribute data set. However, the system fails to improve the extrapolation and interpolation ability and also the accuracy of the system is lower. With an intention to increase the accuracy and to improve the performance through achieving much better extrapolation and interpolation of kernels, in this research, three approaches are proposed namely RBF kernel using KL divergence, Polynomial kernel using probability distribution and finally Piece level combination in mixed kernels. A better accuracy could thus be ensured using these techniques. Consequently this will increase the performance and efficiency of the system. The proposed system is fast, efficient, easily implemented and widely used approach of missed data estimation in mixed attributes.

References

[1] 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. http://dx.doi.org/10.1080/713827181

[2] K. Lakshminarayan et al., "Imputation of Missing Data in Industrial Databases," Applied Intelligence, vol. 11, pp. 259-275, 1999. http://dx.doi.org/10.1023/A:1008334909089

[3] Zhang, S.C., Qin, Y.S., Zhang, J.L., Zhu, X.F., Zhang, C.Q. (2008). Missing Value Imputation Based on Data Clustering. *Transactions on Computational Science Journal*, LNCS 4750, pp 128-138.

[4] Qin, Z.X. (2007). *Multiple costs and their combination in cost-sensitive learning*. PhD Thesis, University of Technology Sydney, 2007.

[5] 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.

[6] 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. http://dx.doi.org/10.1109/TKDE.2005.188

[7] Y.S. Qin et al., "Semi-Parametric Optimization for Missing Data Imputation," Applied Intelligence, vol. 21, no. 1, pp. 79-88, 2007. http://dx.doi.org/10.1007/s10489-006-0032-0

[8] R. Little and D. Rubin, Statistical Analysis with Missing Data, second ed. John Wiley and Sons, 2002. http://dx.doi.org/10.1002/9781119013563

[9] U. Dick et al., "Learning from Incomplete Data with Infinite Imputation," Proc. Int'l Conf. Machine Learning (ICML '08), pp. 232- 239, 2008. http://dx.doi.org/10.1145/1390156.1390186

[10] M. Huisman, "Missing Data in Social Network," Proc. Int'l Sunbelt Social Network Conf. (Sunbelt XXVII), 2007.

[11] Q.H. Wang and R. Rao, "Empirical Likelihood-Based Inference under Imputation for Missing Response Data," Annals of Statistics, vol. 30, pp. 896-924, 2002. http://dx.doi.org/10.1214/aos/1028674845

[12] 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. http://dx.doi.org/10.1016/j.eswa.2008.01.059

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

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

[15] 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. http://dx.doi.org/10.1016/S0304-4076(03)00157-X

[16] V.C. Raykar and R. DuraiswamiFast, "Fast Optimal Bandwidth Selection for Kernel Density Estimation," Proc. SIAM Int'l Conf. Data Mining (SDM '06), pp. 524-528, 2006.