ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Higher Order Kernel Function Algorithm for Imputing Missing Values

Ganga.A.R*
P.G.Scholar
Dept. of Computer Science and Engineering
Anna University of Technology Coimbatore, India
gangaharipriya@gmail.com

B.Lakshmipathi
Assistant Professor
Dept. of Computer Science and Engineering
Anna University of Technology Coimbatore, India
lkpathi_2004@yahoo.com

Abstract: Many types of experimental data are with missing values that may occur for a variety of reasons. Most of the data analyses such as classification methods, clustering methods and dimension reduction procedures require complete data. Hence researchers must either remove missing data or preferably estimate the missing values before such procedures are applied. Missing data imputation is a key issue in learning from incomplete data. Various techniques have been developed to deal with missing values in data sets with homogenous attributes. But those approaches are independent of either continuous or discrete values. Recently a new proposal came for setting of missing data imputation with heterogeneous attributes thus by contributing for both continuous and discrete data. This study proposes a higher order spherical kernel based iterative estimator to impute mixed-attribute data sets. Spherical kernel based estimator will give better results than other estimators.

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

I. INTRODUCTION

Many types of experimental data, especially expression data obtained from microarray experiments and air pollutant data obtained from air sample collecting machine are frequently peppered with missing values (MVs) that may occur for a variety of reasons. Because many data analyses such as classification methods, clustering methods and dimension reduction procedures require complete data, researchers must either remove the data with MVs, or, preferably, estimate the MVs before such procedures can be employed. Consequently, many algorithms have been developed to accurately impute missing values.

Missing data imputation[1] aims at providing estimations for missing values by reasoning from observed data. Because missing values can result in bias that impacts on the quality of learned patterns or/and the performance of classifications, missing data imputation has been a key issue in learning from incomplete data. Various techniques have been developed with great successes on dealing with missing values in data sets with homogeneous attributes (their independent attributes are all either continuous or discrete). However, these imputation algorithms cannot be applied to many real data sets, such as equipment maintenance databases, industrial data sets, and gene databases, because these data sets are often with both continuous and discrete independent attributes. 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 studies a new setting of missing data imputation, i.e., imputing missing data in mixed-attribute data sets using kernel functions.

Imputing mixed-attribute data sets can be taken as a recent problem in missing data imputation because only few estimators designed for imputing missing data in mixed attribute data sets. The challenging issues include how to measure the relationship between instances (transactions) in a mixed-attribute data set, and how to construct hybrid estimators using the observed data in the data set. A nonparametric iterative imputation method based on a mixture kernel is the solution. It first constructs a kernel estimator to infer the probability density for independent attributes in a mixed-attribute data set. And then, a mixture of kernel functions (a linear combination of two single kernel functions, called mixture kernel) is designed for the estimator in which the mixture kernel is used to replace the single kernel function in traditional kernel estimators. These estimators are referred to as mixture kernel estimators. Based on this, two consistent kernel estimators are constructed for discrete and continuous missing target values, respectively, for mixed-attribute data sets. Further, a mixture-kernel-based iterative estimator is proposed to utilize all the available observed information, including observed information in incomplete instances (with missing values. These experiments were conducted on UCI data sets at different missing ratios.

Imputation is the substitution of some value for a missing data point or a missing component of a data point. Once all missing values have been imputed, the dataset can then be analysed using standard techniques for complete data. The analysis should ideally take into account that there is a greater degree of uncertainty than if the imputed values had actually been observed, however, and this generally requires some modification of the standard complete-data analysis methods. Many imputation techniques are available. Hot- deck imputation and cold-deck imputation are the common techniques. Imputation is not the only method available for handling missing data. It usually gives better results than 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 in many circumstances. The expectation-maximization algorithm is a method for finding maximum likelihood estimates that has been widely applied to missing data problems. Other successful methods include computational intelligence methods.

II. RELATED WORK

G. Batista and M. Monard [2] proposed an analysis of four missing data treatment methods for supervised learning.

The four 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. The experiments evaluated the efficiency of the k-nearest neighbour algorithm as an imputation method to treat missing data, comparing its performance with the performance obtained by the internal algorithms used by C4.5 and CN2 to learn with missing data, and by the mean or mode imputation method. R. Caruana [3] proposed a non-parametric EM-style algorithm for imputing missing values. 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 main feature is that E and M steps collapse 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 main advantages are that (1) it is more efficient for moderate size datasets and (2) it is less susceptible to errors. U.Dick et al [4] proposed a learning method from incomplete data with infinite imputation. They derived a generic joint optimization problem in which the distribution governing the missing values is a free parameter. The optimal solution concentrates the density mass on finitely many imputations, and provides a corresponding algorithm for learning from incomplete data. The instantiations of the general learning method consistently outperform single imputations. Z. Ghahrami and M. Jordan[5] proposed mixture models for learning from incomplete data. They aimed at two things- to place current neural network approaches to missing data within a statistical framework and to describe a set of algorithms that can handle clustering, classification and function approximation from incomplete data in an efficient manner.

The algorithms are based on mixture modelling and make two appeals to the Expectation-Maximization (EM) principle both for the estimation of mixture components and for coping with missing data. M. Huisman [6] studied about imputation of missing network data and some simple procedures are proposed. Imputing the unconditional mean, Imputation by Reconstruction, Imputation using preferential attachment and Hot deck imputation are the different approaches. Performances of different techniques are also evaluated. Effects of missing data and effects of imputation are also verified. By simulations 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. He proposed definitions for irrelevance and for two degrees of relevance. This improved the understanding behaviour of previous subset selection algorithms. A method for feature subset selection using cross-validation that can be applied on any induction algorithm is proposed. R. Marco [8] proposed a method of learning Bayesian networks from incomplete data. A new deterministic method to learn the graphical structure of a BBN from an incomplete database is introduced.

This method, when coupled with a method able to assess the parameters of a given a graphical model from an incomplete database, give rise to systems that are able to extract a complete Bayesian Network from an incomplete database. Significant features are robustness and independence of its execution time from the number of missing data. J.R. Quinlan [9] compared the effectiveness of different approaches used for the classification of attributes with unknown values. Different approaches handled are unknown values when partitioning, unknown values when classifying and unknown values in selecting tests. Different results for each approach are also suggested. J.Racine and Q.Li [10] proposed nonparametric estimation of regression functions with both categorical and continuous data. A data driven method of bandwidth selection is also proposed. The new estimator performs much better than 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.

The algorithm is for univariate Gaussian kernel based density derivative estimation that reduces the computational complexity. The speedup achieved in this problem is demonstrated using the solve-the-equation plug-in method. G.F. Smits and E.M. Jordaan [12] gave an improved SVM regression using mixtures of kernels. Kernels are used in Support Vector Machines to map the learning data into a higher dimensional feature space. To overcome disadvantages of single kernels mixtures of kernels can be used which can give additional abilities. 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. That is, there is a known cost associated with each attribute (variable or test) when obtaining its values. Cost-sensitive learning algorithms should make use of only known values.

The method produces decision trees with minimal total cost of tests and misclassifications on the training data. S.C. Zhang [14] studied about parimputation ie, from imputation and null- imputation to partially imputation. Some missing data are imputed when there are some complete data in a small neighborhood of the missing data and, other missing data without imputation are given up in applications, such as data mining and machine learning. Missing data mechanisms are classified into three categories as Missing Completely at Random (MCAR),Missing at Random (MAR) and Nonignorable. In practice it is usually difficult to meet the nonignorable assumption. MAR is an assumption that is more often (MCARm is a special case of MAR), but not always tenable. The more relevant and related predictors one can include in statistical models, the more likely it is that the MAR assumption will be met. W. Zhang [15] proposed an association based multiple imputation technique in multivariate datasets. A framework of association rules is applied to multiple imputations in multivariate datasets with missing numeric and categoric data. Association relationships among variables and statistical features of possible variable values are combined to predict the most possible values of missing data.

III. IMPUTATION AND KERNEL METHODS

Imputing mixed-attribute data sets can be taken as a new problem in missing data. Commonly used methods to impute missing values include parametric and nonparametric regression imputation methods.

The parametric method, such as linear regression is superior while the data sets are adequately modelled.

However, in real applications, it is often impossible to know the distribution of the data set. Therefore, the parametric estimators can lead to highly bias, and the optimal control factor settings may be miscalculated. For this case, nonparametric imputation method[10] can provide superior fits by capturing the structure of the data set.

However, these imputation methods are designed for either continuous or discrete independent attributes. For example, the well-established imputation methods are developed for only continuous attributes. And these estimators cannot handle discrete attributes well. Some methods, such as C4.5 algorithm[3], association-rule-based method, and rough-set-based method, are designed to deal with only discrete attributes. In these algorithms, continuous attributes are always discretized before imputing. This possibly leads to a loss of useful characteristics of the continuous attributes. There are some conventional imputation approaches designed for discrete attributes using a "frequency estimator" in which a data set is separated into several subsets or "cells."

However, when the number of cells is large, observations in each cell may not be enough to nonparametrically estimate the relationship among the continuous attributes in the cell.

When facing with mixed independent attributes, some imputation methods take the discrete attributes as continuous ones, or other methods are used. Some reports, for instance and selected to smooth the mixed regressors, but without taking the selection of bandwidth into account. Therefore, Racine and Li[10] proposed a natural extension of the method to model the settings of discrete and continuous independent attributes in a fully nonparametric regression framework. To find missing values from incomplete data Expectation maximization algorithm was proposed [16]. The importance of missing value estimation is studied in detail in [17]. Recently Zhu, Zhang and Jin[18] proposed a method for estimation of missing values for heterogeneous attribute data sets. They used a mixture kernel function for the imputation. A combination of polynomial and RBF kernel is used. They proposed an algorithm for this missing value estimation.

Kernel methods (KMs) are a class of algorithms for pattern analysis. The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in general types of data (such as sequences, text documents, sets of points, vectors, images, etc.).Kernel methods approach the problem by mapping the data into a high dimensional feature space, where each coordinate corresponds to one feature of the data items, transforming the data into a set of points in a Euclidean space. In that space, a variety of methods can be used to find relations in the data. Since the mapping can be quite general (not necessarily linear, for example), the relations found in this way are accordingly very general. This approach is called the kernel trick.

KMs owe their name to the use of kernel functions, which enable them to operate in the feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors.

Commonly using kernel functions [19] are

Linear Kernel

The Linear kernel is the simplest kernel function. It is given by the inner product <x,y> plus an optional constant **C**. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts, i.e. KPCA with linear kernel is the same as standard PCA.T is also a constant.

$$K(x,y)= x^T y + C \quad (1)$$

Polynomial Kernel

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

$$K(x.y) = (\alpha x^T + c)^d \quad (2)$$

Adjustable parameters are the slope **alpha**, the constant term **c** and the polynomial degree **d**.

Radial Basis Function Kernel

The general radial basis function kernel is given by

$$K(x,y) = \exp(-(x-y)^2 / \sigma^2) \quad (3)$$

where σ is an adjustable constant.

Gaussian Kernel

The Gaussian kernel is an example of radial basis function kernel.

$$K(x,y)= \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (4)$$

Alternatively, it could also be implemented using

$$K(x,y)= \exp\left(-\gamma\|x-y\|^2\right) \quad (5)$$

The adjustable parameter **sigma** plays a major role in the performance of the kernel, and should be carefully tuned to the problem at hand. If overestimated, the exponential will behave almost linearly and the higher-dimensional projection will start to lose its non-linear power. In the other hand, if underestimated, the function will lack regularization and the decision boundary will be highly sensitive to noise in training data.

Exponential Kernel

The exponential kernel is closely related to the Gaussian kernel with sigma as constant, with only the square of the norm left out. It is also a radial basis function kernel.

$$K(x,y)= \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (6)$$

Laplacian Kernel

The Laplace Kernel is completely equivalent to the exponential kernel, except for being less sensitive for changes in the sigma parameter. Being equivalent, it is also a radial basis function kernel.

$$K(x,y)= \exp\left(-\frac{\|x-y\|}{\sigma}\right) \quad (7)$$

It is important to note that the observations made about the sigma parameter for the Gaussian kernel also apply to the Exponential and Laplacian kernels.

Spherical Kernel

The spherical kernel is a higher order kernel with order 3 and sigma is the adjustable parameter.

$$K(x,y) = 1 - \frac{3}{2}\frac{\|x-y\|}{\sigma} + \frac{1}{2}\left(\frac{\|x-y\|}{\sigma}\right)^3 \quad (8)$$

IV. MIXED KERNEL USING HIGHER ORDER KERNEL

A mixed kernel approach for estimation of missing values was proposed by Zhu etal[17]. Even though the

existing system presented the imputation of the missing values for heterogeneous attributes, it has some drawbacks. In the research of the missing value imputation, the existing systems are not well defined for the different types of data that needs better interpolation and extrapolation.

a. ***Interpolation:*** Interpolation is a method of constructing new data points within the range of a discrete set of known data points. In many applications one often has a number of data points, obtained by sampling or experimentation, which represent the values of a function for a limited number of values of the independent variable. It is often required to interpolate (i.e. estimate) the value of that function for an intermediate value of the independent variable. This may be achieved by curve fitting or regression analysis. It can be considered as the ability to learn from the data

b. ***Extrapolation:*** Extrapolation is the process of constructing new data points. It is similar to the process of interpolation, which constructs new points between known points, but the results of extrapolations are often less meaningful, and are subject to greater uncertainty. It may also mean extension of a method, assuming similar methods will be applicable. It can be considered as the ability to predict unseen data.

The main issue here addressed is how to impute the missing values in a mixed-attribute data set in an efficient way. Initially a nonparametric iterative imputation method is presented. In this a kernel functions for the discrete attributes are studied and then a mixture kernel function is proposed by combining a discrete kernel function with a continuous one. Further an estimator is constructed based on the mixture kernel. Finally the nonparametric iterative imputation algorithm extended from a single kernel to a mixture of kernel is designed and analyzed. Then the proposed system is evaluated and compared with other approaches. Instead of the existing system this study proposes a mixture kernel based on the spherical kernel function so as to achieve better interpolation and extrapolation. Here for the experimental studies a set of data from the real applications and datasets from the UCI repository are used. From experimental backgrounds it was found that the proposed approach will work better than the existing system. Using higher order kernel functions will absolutely give better results. For the combination for making mixture kernel RBF or any other kernel can be used.

A. *Algorithm Design:*

The algorithm is designed as follows:

Mixture Kernel approach based on combinations of RBF kernel given by (3) and spherical kernel given by (8) gives results better than previous approaches. Different parameters considered during the experimental study are the different values given to x any y and the constant σ. The experiment repeats for different values of σ.

All the imputed values are used to impute the subsequent missing values, i.e, the (t+1)th iteration imputation is carried out based on the imputed results of the tth imputation, until the filled-in values converge or begin to cycle or satisfy the demands of the users. This is the main advantage of this particular algorithm

In the first iteration of imputation, roughly all the missing values are imputed. Since the second iteration of imputation, each of the iteration- imputation is carried out based on former imputed results with the non- parametric kernel estimator. During the imputation process when missing value of x is imputed, all other missing values are regarded as observed values. The iteration imputation for missing continuous attributes will be terminated when the filled-in values converge or begin to cycle.

B. *Experimental Study:*

For the experimental studies data sets from UCI repository are taken. Data sets having mixed attributes along with missing values are chosen. Most of the experiments are done in Automobile data set with 26 attributes having discrete, continuous and mixed values.

Table I. Automobile Database used in Experiment

Data Set Characteristics	**Multivariate**	**Number of Instances**	**205**
Attribute Characteristics	Categorical, Integer, Real	Number of Attributes	26
Associated Tasks	Regression	Missing Values?	Yes

Firstly an algorithm was developed to find missing values for discrete attributes using polynomial kernel given by (2). Missing values for different attributes with discrete values only are found using the algorithm. To systematically study the performance of the algorithm, known values are made missing and then values are found using the algorithm. Secondly similar algorithm using spherical kernel given by (8) is developed. Missing values for same attributes are found out. Spherical kernel method having order three give more accurate results. Next Radial Basis Kernel method based on (3) is developed for finding missing values for continuous attributes. As next step, for attributes with mixed values i.e. having both continuous and discrete values, a mixture kernel method using polynomial and rbf kernel method is developed and thereafter combination of spherical and rbf kernels are developed.

The different parameters considered during experiment are the order of the kernel functions using, the constant sigma and different inputs to the kernel functions. Random values are selected in each run of the algorithm.

A chart showing the accuracy of true value to calculated values for both mixed kernel approaches i.e. combination of polynomial and rbf as well as spherical and rbf is shown below: Fig 1 gives the accuracy of the two values based on mixed kernel using polynomial and rbf kernel approaches. Fig 2 shows same for spherical and rbf approach. From the charts it is clear that second approach gives better results than first.

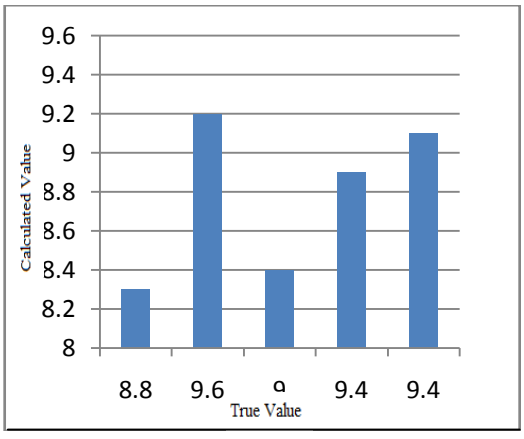

Figure 1: Mixed Kernel 1

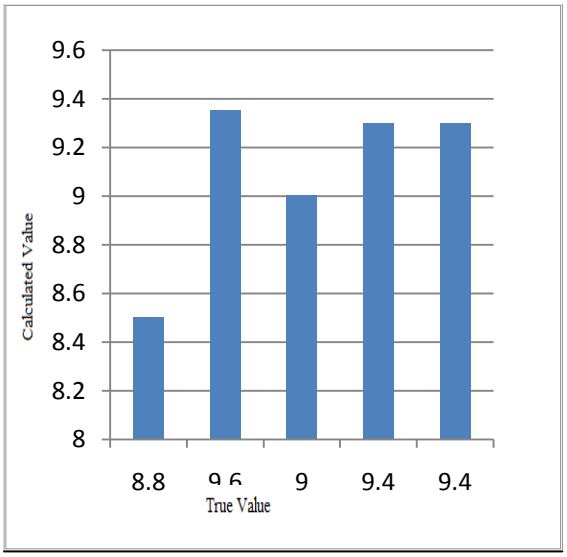

Figure 2: Mixed Kernel 2

V. CONCLUSIONS AND FUTURE WORK

A mixture kernel-based iterative nonparametric estimator based on higher order kernel functions for data sets having both continuous and discrete independent attributes is designed. It utilizes all available observed information, including observed information in incomplete instances (with missing values), to impute missing values, whereas existing imputation methods use only the observed information in complete instances (without missing values).

In future, this work can be extended to check with different combinations of kernel methods to achieve better results. Also the kernel based approach can be extended to find missing values for heterogeneous attribute data sets.

VI. 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