Journal of Theoretical and Applied Information Technology
10th February 2015. Vol.72 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MISSING VALUE IMPUTATION USING FUZZY POSSIBILISTIC C MEANS OPTIMIZED WITH SUPPORT VECTOR REGRESSION AND GENETIC ALGORITHM

[1]P.SARAVANAN, [2]P.SAILAKSHMI

[1]Asst. Prof., Department of Information Technology, Sathyabama University, Chennai, INDIA

[2]Faculty, Vista Mind Pvt. Ltd., Chennai, INDIA

E-mail: psaravanan76@gmail.com

ABSTRACT

Quality data mining results can be obtained only with high quality input data. So missing data in data sets should be estimated to increase data quality. Here comes the importance of efficient methods for imputation of missing values. If the values are Missing At Random (MAR), it can be estimated using some complex manner from available data. For such an estimation of values, a combination of fuzzy c means and possibilistic c means algorithms are used in the proposed system. Thus combining the advantages of fuzzy c means algorithm, such as data can belongs to more than one cluster which gives best result for overlapped data etc and that of possibilistic c means such as handling noisy data effectively. Proposed system considers both membership function and typicality of the data. Fuzzy-Possibilistic c means method is optimized using Genetic Algorithm with Support Vector Regression (SVRGA). The main purpose of SVRGA is to minimize the error. Support Vector Regression model must be trained with complete records. Genetic Algorithm is used to select new parameters from existing population. If the error is found to be minimum then it is assumed that parameters are optimized and the dataset does not contain incomplete records. If the error is not minimum again estimate the missing values using fuzzy possibilistic c means clustering with new parameters. The system is tested with two different real time datasets, Iris and marine db with various standard missing ratios. The performance of proposed method is calculated using Random Mean Square Error (RMSE) and compared with competitor. The graphs show the system proposed in this work is performing well.

Keywords: *Missing value Imputation, Fuzzy Possibilistic C Means, Support Vector Regression, Genetic Algorithm, Multiple Imputations.*

1. INTRODUCTION

The amount of data stored in files and databases is increasing in an exponential rate due to automated data collection equipments. At the same time, sophisticated information from this huge data is very small. Extracting hidden information in such huge data base is often called as data mining. To ensure the usefulness and accuracy of the results in data mining process, data should be correct, complete and consistent. As per survey, 80% of the effort is used for data preparation, which makes the data as correct, complete and consistent. Even though Data processing includes many types includes data cleaning, integration etc. Data cleaning is an important task which makes data complete and error free.

Missing values are highly undesirable in data mining, machine learning and other information systems. In recent years, much research regarding missing value estimation and imputation has been performed. To deal with missing values in datasets: ignoring, deleting, zero or mean estimation methods might be used instead of imputation methods. However, the primary disadvantages of these estimation methods are the loss of efficiency due to discarding incomplete observations and biases in estimates when data are missing in a systematic manner. These disadvantages reduce data quality. Missing values should be estimated to increase data quality.

2. LITERATURE REVIEW

Ibrahim Berkan Aydilek, AhmetArslan [1] proposed a hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Although the system imputes the missing values it fails to deal noisy data effectively. Marie-Jeanne Lesot [2] proposed Typicality-based clustering.

Typicality degrees are defined to build prototypes that characterize data subcategories, taking into account both the common points of the category members and their distinctive features as compared to other categories. The System lacks in performance with overlapped data. Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek [3] proposed a Possibilistic Fuzzy c-Means Clustering Algorithm that generated both membership and typicality values while clustering the unlabeled data. FPCM constraints the typicality values, so that the sum over all data points of topicalities to a cluster is one. I.e., row sum constraint, which makes the system difficult to handle huge datasets. Nasraoui and R. Krishnapuram [5] suggested Crisp interpretations of fuzzy and possibilistic clustering algorithm. In their work, they derive equivalent crisp objective functions for each of the three families of algorithms. It is found that the concept of memberships can be totally bypassed using these crisp reformulations. Since system does no optimization, it is sensitive to initial guesses. Gil, E.G. Sarabia, J.R. Llata, J.P. Oria [6] put forward Improved Fuzzy c-means clustering for noise reduction, enhancement and reconstruction of 3D ultrasonic images. A possibilistic approach was proposed in a previous paper for C-means clustering, and two algorithms realizing this approach were reported in two previous works mentioned above. Although the possibilistic approach is sound, these two algorithms tend to find identical clusters. D.E. Goldberg [7] proposed Sizing populations for serial and parallel genetic algorithms. He considers the effect of stochasticity on the quality of convergence of genetic algorithms (GAs). Here premature convergence results in suboptimal population. Li, H. Gu, L.Y. Zhang [13] proposed a fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. Since the proposed system does no optimization, it is sensitive to initial guesses. C.H. Huang, H.Y. Kao [11], studied Interval regression analysis with soft-margin reduced support vector machine. The Support Vector Machine (SVM) has shown to be an efficient approach for a variety of classification problems. This requires less computational time, but does not yield sensible results for outlier values. Zaifei Liao, Xinjie Lu, Tian Yang, Hongan Wang [14], suggested a fuzzy K-means clustering algorithm over sliding window. Fuzzy set theory is motivated by the practical needs to manage and process uncertainty inherent in real world problem solving. It is useful in applications to data mining, conflict analysis, and so on. But system does not deal outliers effectively. The system proposed in this paper takes all the pros and cons of the above mentioned work and optimized.

2.1 Missing Data

Missing values occurs in the dataset due to many reasons. Perhaps the field was not applicable, the event did not happen or the data was not available. It could be that the person who entered the data did not know the right value, or did not care if a field was not filled in.

However, there are many data mining scenarios in which missing values provide important information. The meaning of the missing values depends largely on context. For example, a missing value for the date in a list of invoices has a meaning substantially different from the lack of a date in column that indicates an employee hire date. There are three types of data described in the literature.

1. Missing completely at random (MCAR) – The missing value has no dependency on any other variable.
2. Missing at random (MAR) – The missing value depends on other variables. The missing value can be estimated using other variables.
3. Missing not at random (MNAT) – The missing value depends on other missing values, and thus missing data cannot be estimated from existing variables.

In this paper, we are considering only MAR values that are deducible in some complex manner from the remaining data. Training SVR is not possible if the values are not missing at random. In Table 1, we present a section of a dataset with missing values. In this paper, we aim to estimate missing values using fuzzy possibilistic c-means optimized with support vector regression and a genetic algorithm. Y1, Y2, Y3, Y4, Y5 and Y6 are records (rows). X1, X2, X3, X4 and X5 are attributes (columns). Y2, Y5 and Y6, which do not have any missing values, are 'complete' rows, and Y1, Y3 and Y4, which have missing values, are called 'incomplete' rows.

2.2 Fuzzy Possibilistic C Means Algorithm

Estimation of values is done by Fuzzy Possibilistic c means clustering algorithm. It is a combination of Fuzzy c means and Possibilistic c means algorithm for clustering. Thus combining the advantages of fuzzy c means algorithm such as data can belongs to more than one cluster, gives best result for overlapped data etc. and that of possibilistic c means such as handling noisy data effectively. Proposed system considers both membership function and typicality of the data, which helps to reduce the error rate. Membership function is the degree how much an item belongs to one cluster. So each item should have one membership value for each cluster.

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}}$$

Where

$$d_{ji} = \sqrt{\Sigma\ (x_i - c_j)^2}$$

Typicality is the degree which tells us how much a cluster holds one item. So each cluster should have one typicality value for each item. Typicality degrees are defined to build prototypes that characterize data subcategories, taking into account both the common points of the category members and their distinctive features as compared to other categories. The typicality of a point for its category depends on two complementary notions, called internal resemblance and external dissimilarity: the former is defined as the point resemblance to the other members of the category, the latter as its dissimilarity from members of the category.

Typicality step: for each point x ∈ X and each cluster Cr, r = 1.. c

(a) Compute the internal resemblance R(x,Cr)=avg((x, y), y < Cr)

(b) Compute the external dissimilarity D(x,Cr)=avg((x, y), y <Cr)

(c) Compute the typicality degree T(x,Cr)='(R(x,Cr), D(x,Cr))

Assignment step: for each point x ∈ X

(a) If x is typical for no cluster, i.e. maxr T(x,Cr) < 0.1, assign x to a fictitious cluster, C0.

(b) Else if x typicality is not clear, i.e. T1(x) − T2(x) < 0.02, where Ti(x) is the i-th biggest value of T(x, Cr), r = 1..c, assign x to the fictitious cluster C0.

(c) Else assign x according to the maximal typicality degree, i.e. to Cr where r = argmaxs T(x, Cs).

3. FPCM-SVGRA IMPUTATION MODEL

The architecture of the proposed system is shown in Figure 1. The proposed model classifies the given dataset into two clusters. One with complete records and other with missing values which will be supplied as input to fuzzy possibilistic estimator. The fuzzy possibilistic module clusters the data set using FPCM algorithm which is later trained by Support Vector Regression. A non-linear model given by SVM is trained to compute the missing values. The computed values are taken for accuracy calculation using Root Mean Square Error (RMSE) method. Once the RMSE value is within the threshold, then the model trained by SVR gets used to make the dataset as complete. Parameters gets optimized using genetic algorithm and the entire process gets iterated till the list of parameters got optimized.

3.1 Algorithm Used

The algorithm proposed in this paper possesses the following steps.

1. Load Dataset to database.
2. Create another similar database named incomplete and complete.
3. Split given data into two dataset and load into databases
4. Calculate cluster number(c)
5. Compute the weighting factor 'm' for every cluster.
6. Get the cluster center c_j for every cluster.
7. Calculate the distance d
8. Compute the value for membership function
9. Calculate the typicality
10. Calculate the value for missing cells
11. Build SVR model by training with known data.
12. Train the support vector regression algorithm with complete rows
13. Compare the fuzzy possibilistic c-means output with the SVR output
14. The minimized error function can be computed as error = $(X– Y)^2$ where X is the output of the support vector regression (SVR) prediction and Y is the output of fuzzy possibilistic c-means algorithm prediction
15. Obtain the optimized c and m parameters by using the genetic algorithm to minimize the difference between the SVR output and the fuzzy possibilistic c-means output.
16. Estimate the missing values using fuzzy possibilistic c-means with optimized parameters.

4. EXPERIMENTAL IMPLEMENTATION

We used two datasets that are frequently used in literature and the UCI Repository of Machine Learning Databases. Table 2 describes the abstract properties of dataset Iris and Marine DB used in this work. Testing on more datasets and working with several numbers and types of missing values are needed to determine how the algorithm generalizes. All datasets are artificially regenerated such that they have 5%, 10%, 15%, and 20% missing value ratios.

Table 2: Datasets Used For The Experiment

Data Set Name	Records	Attributes
Iris	150	5
Marine DB	700	7

The proposed method is coded using java, and the LOGIT toolbox was used for SVR. The genetic algorithms are implemented by generating a population and creating a new population by performing the following procedures: reproduction, crossover, and mutation used a population size of 5, 7 generations, a crossover fraction of 60% and a mutation fraction of 3%.

5. RESULTS AND DISCUSSION

The efficiency of the missing data estimation system is evaluated using the root mean standard error (RMSE). The root mean standard error measures the error between the real values and the estimated values and quantifies the accuracy of the prediction.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_i)^2}{N}}$$

Following Figures shows the performance of the proposed system for the two standard datasets, Marine DB datasets, for various standard missing ratios 5%, 10%, 15% and 20%. The results show that the proposed method (FPCM-SVRGA) is performing well compare to its contrasting method (FCM-SVRGA) in missing value estimation.

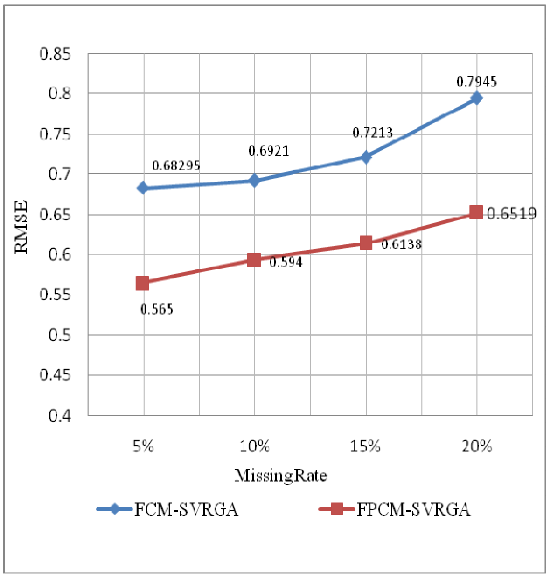

Figure 2: Performance Analysis For IRIS Dataset

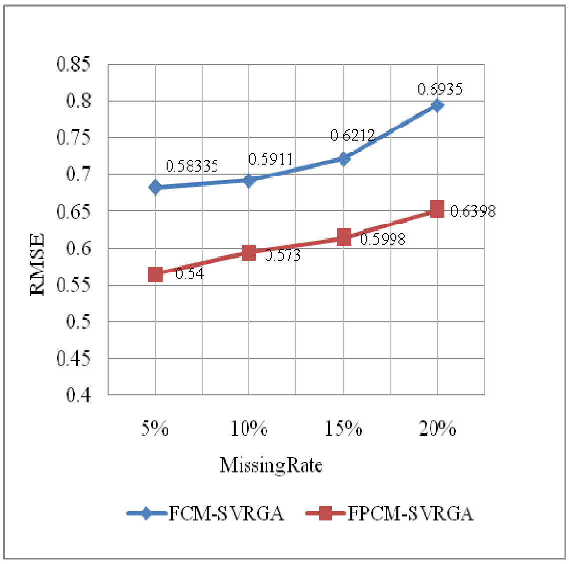

Figure 3: Performance Analysis For Marine DB Dataset

6. CONCLUSION

In this paper, a hybrid method that uses a reliable machine learning technique known as support vector regression and a genetic algorithm was used with fuzzy possibilistic clustering to estimate missing values. Complete data were clustered based on their similarity, and both fuzzy and possibilistic principles were used during clustering. Therefore, each missing value becomes a member of more than one cluster centroids, which yields more sensible imputation results. Two datasets with different characteristics were used in this paper, and the cluster size and the weighting factor parameters are optimized according to the corresponding dataset. Better imputation accuracy is achieved by FPCM-SVRGA as compared to the FCM-SVRGA method. The experimental results performance demonstrated that the Fuzzy Possibilistic C-Means SVRGA imputation yields a more sufficient, sensible estimation accuracy ratio for suitable clustering data.

REFERENCES:

[1] Ibrahim Berkan Aydilek , AhmetArslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm", *Information Sciences*, Vol. 233, June 2013, pp. 25–35.

[2] Marie-Jeanne Lesot, "Typicality based clustering", *International Journal of Information Technology and Intelligent Computing,* Vol. 1, 2006, pp. 279-292.

[3] James C. Bezdek, James Keller, Raghu Krishnapuram and Nikhil R. Pal, "Fuzzy Models and Algorithms for Pattern Recognition and Image Processing", *Kluwer Academic Publishers*, TA 1650, 1999, pp89.

[4] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering", *IEEE Transactions on Fuzzy Systems*, Vol. 1, May 1999, pp. 98-110.

[5] Olfa Nasraoui, Maha Soliman, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", *IEEE Transactions on Knowledge & Data Engineering*, Vol. 20, February 2008.

[6] M. Gil, E.G. Sarabia, J.R. Llata, J.P. Oria, "Fuzzy c-means clustering for noise reduction, enhancement and reconstruction of 3D ultrasonic images", *Proceedings of the Emerging Technologies and Factory Automation*, Vol. 1, 1999, pp. 465-472.

[7] D.E. Goldberg, "Sizing populations for serial and parallel genetic algorithms", *Proceedings of the Third International Conference on Genetic Algorithms*, 1989.

[8] J. Hajek, Z.k. Sidak, "Theory of Rank Tests", *Academic Press, New York*, 1967.

[9] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", *Academic Press*, 2001.

[10] R.J. Hathaway, J.C. Bezdek, "Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm", *Pattern Recognition Letter, Science Direct,* Vol. 23, 2002, pp. 151–160.

[11] C.H. Huang, H.Y. Kao, "Interval regression analysis with soft-margin reduced support vector machine", *Proceedings of Next-Generation Applied Intelligence,* Vol. 5579 2009, pp. 826–835.

[12] J. Van Hulse, T.M. Khoshgoftaar, "Incomplete-case nearest neighbor imputation in software measurement data", *Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration*, 2007, pp. 630–637.

[13] D. Li, H. Gu, L.Y. Zhang, "A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data", *Expert System Applications,* Vol. 37, 2010, pp. 6942–6947.

[14] Zaifei Liao, Xinjie Lu, Tian Yang, Hongan Wang, "Missing data imputation: A fuzzy K-means clustering algorithm over sliding window", *Fuzzy Systems and Knowledge Discovery,* Vol. 3, 2009, pp. 133–137.

[15] C.P. Lim, J.H. Leong, M.M. Kuan, "A hybrid neural network system for pattern classification tasks with missing features", *IEEE Transactions Pattern Analysis*, Vol. 27, 2005, pp.648-653.

[16] R.J.A. Little, D.B. Rubin, "Statistical Analysis with Missing Data", *Wiley*, New York, 1987.

[17] M.S. Mahmoud, M.F. Emzir, "State estimation with asynchronous multi-rate multi-smart sensors", *Information Sciences,* Vol. 196, 2012 pp. 15–27.

[18] T. Marwala, "Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques", *Information Science Reference*, *Hershey PA*, 2009.

[19] T. Marwala, S. Chakraverty, "Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm", *Current Science, India*, Vol. 90, 2006, pp. 542–548.

[20] Z.Q. Meng, Z.Z. Shi, "Extended rough set-based attribute reduction in inconsistent incomplete decision systems", *Information Sciences*, Vol. 204, 2012, pp. 44–69.

[21] S. Mohamed, T Marwala, "Neural network based techniques for estimating missing data in databases", *16th Annual Symposium of the Pattern Recognition Association of South Africa,* Langebaan, 2005, pp. 27–32.

[22] F.V. Nelwamondo, S. Mohamed, T. Marwala, "Missing data: a comparison of neural expectation maximization techniques", *Cornell University*, Vol. 93, 2007, pp. 1514–1521.

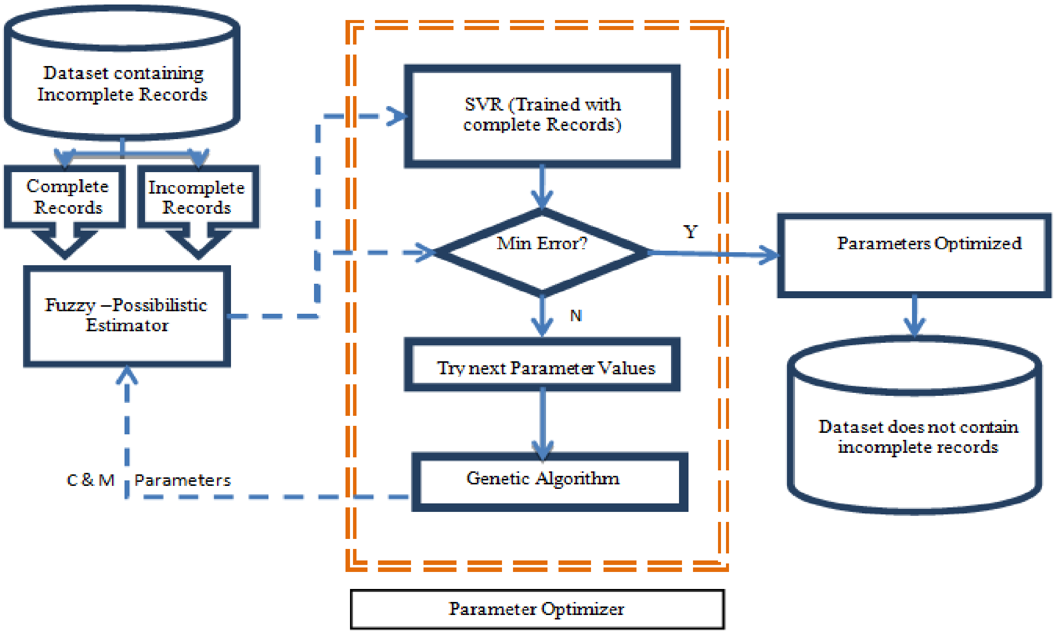
Dataset containing Incomplete Records
Complete Records
Incomplete Records
Fuzzy –Possibilistic Estimator
C & M
Parameters
SVR (Trained with complete Records)
Min Error?
Y
N
Try next Parameter Values
Genetic Algorithm
Parameter Optimizer
Parameters Optimized
Dataset does not contain incomplete records

Figure 1: The Proposed FPCM-SVRGA Imputation Method

Table 1: Database Containing Missing Values

	X1	X2	X3	X4	X5
Y1	0.113524	0.084785	?	0.625473	0.015023
Y2	0.112537	0.138211	0.159420	0.625473	0.068545
Y3	0.110563	?	0.144928	0.624212	0.083568
Y4	0.110563	0.170732	0.146998	0.623581	?
Y5	0.108588	0.129501	0.144928	0.624212	0.076056
Y6	0.108588	0.082462	0.112836	0.626103	0.015023