Evolving clustering based data imputation

Conference Paper · March 2014

DOI: 10.1109/ICCPCT.2014.7054988

CITATIONS
18

READS
219

2 authors:

Chandan Gautam
Indian Institute of Technology Indore
26 PUBLICATIONS **334** CITATIONS
SEE PROFILE

Vadlamani Ravi
Institute for Development & Research in Banking Technology
115 PUBLICATIONS **4,681** CITATIONS
SEE PROFILE

Evolving Clustering Based Data Imputation

Published in IEEE Explore, Plz cite this paper: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7054988

Chandan Gautam
SCIS, University of Hyderabad, Hyderabad-500046 (AP), India and
Center of Excellence in CRM and Analytics
Institute for Development & Research in Banking Technology, Hyderabad-500057 (AP), India
email: induindu31@gmail.com

Vadlamani Ravi*
Center of Excellence in CRM and Analytics
Institute for Development & Research in Banking Technology
Hyderabad-500057 (AP), India
email: rav_padma@yahoo.com

***Abstract*— Missing data is an inevitable problem in many disciplines. In this paper, we employed an Evolving Clustering Method (ECM) based imputation method and performed sensitivity analysis of the influence of threshold value (*Dthr*) on imputation results over 12 datasets. We experimented on a large range of *Dthr* values from 0.001 to 0.999, in steps of 0.001, in order to see which value of *Dthr* would perform better imputation compared to K-Means+MLP. Thereby, we provided an upper bound for the *Dthr* value in ECM algorithm. Further, we tested the effectiveness of the online clustering based imputation method on 12 datasets under 10-fold cross validation set up. ECM yielded better performance compared to K-Means + Multilayer perceptron hybrid algorithm, appearing in literature. It is due to strong local learning capability of ECM and selection of an optimal *Dthr* value.**

Keywords*— *Imputation, Local Learning, Evolving Clustering Method, Missing Data

I. INTRODUCTION

As always, some cases of multivariate data may contain missing observations and many real world datasets also contain missing data. There are several reasons for missing data to occur like non response to some fields in data collection process by respondents because of privacy reasons, data entry errors, system failure, ambiguity of the survey questions etc. The missing data is a major problem for analysts because the completeness and quality of the data plays a major role in analyzing the available data. Imputation is known as the substitution of a missing data point or a missing component of a data point by a suitable value. Most of the data mining algorithms cannot work with incomplete datasets. So, imputation of missing data became mandatory [1, 2, 3 and 4]. There are numerous popular methods for dealing with missing data like mean substitution [5], regression method [6], hot and cold deck imputation [7], expectation maximization [8], multiple imputation [9] etc. According to Kline [10], the methods used for missing data are deletion, imputation, model based and machine learning methods.

A. Deletion Procedures

The deletion techniques delete the cases that contain missing data. According to Song and Shepperd [11], there are two forms in deletion approach. They are List wise deletion and pair wise deletion. The list wise deletion ignores the cases or records containing missing values. Pair wise deletion method considers each feature separately and missing data ignored for each feature.

* Corresponding author

B. Imputation Procedures

The imputation techniques include regression imputation, hot and cold deck imputation, multiple imputation and mean imputation. The regression equations are computed each time by considering the attribute containing incomplete value as target variable. The advantage of regression method is preserving the variance and covariance of missing data with other variables.

In Hot and cold deck imputation, the missing values are replaced by the closest components that are present in both vectors for each case with a missing value [12]. In mean imputation, the missing values of a variable are replaced by the mean value of all the remaining cases of that variable. In multiple imputation, we can make combined inferences by analyzing N complete datasets after replacing each value N times.

C. Model-based Procedures

The maximum likelihood method and expectation maximization are the methods under model-based procedures. The maximum likelihood approach assumes that the observed data are a sample drawn from a multivariate normal distribution and the parameters are estimated based on available data and the missing values are determined based on these parameters [13].

D. Machine Learning Methods

These include multi-layer perceptron (MLP), K-Nearest Neighbor (K-NN) and other such methods.

The remainder of this paper is organized as follows: a brief review of literature survey is presented in section II. Section III briefly describes an Evolving clustering method (ECM). An ECM-Imputation method is explained in section IV. Experimental setup is described in section V. The description of the dataset is presented in section VI. Results and discussions are presented in section VII followed by the conclusion in section VIII.

II. LITERATURE SURVEY

Data imputation techniques are categorized into deletion, imputation, model-based and machine learning or soft computing procedures. The machine learning based methods include K-Nearest Neighbor [14], SOM [15], fuzzy-neural network [16], multi-layer perceptron [17], auto-associative

neural network imputation with genetic algorithms [2] etc. Jerez et al. [18], Batista and Monard [14, 19] used K-NN for imputing missing data. Liu and Zhang [20] developed mutual K-nearest neighbor algorithm for classifying incomplete and noisy data. Samad and Harp [21] implemented SOM approach for dealing the missing data. Austin and Escobar [22] applied Monte Carlo simulations to testify the performance of three Bayesian methods that imputed missing data by replacing a simple prior distribution upon the variable that included missing values. In the neural network approach, MLP should be trained as regression model by using the complete records and choosing one variable as target each time.

Various researchers Nordbotten [23], Sharpe and Solly [24], Yoon and Lee [25], Gupta and Lam [26], Nkuna and Odiyo [27] and Silva-Ramirez et al. [28] used MLP for missing data imputation. When auto-associative neural network (AANN) is used for imputation, the network is trained for predicting the inputs by taking the same input variable as a target [29, 30]. Ragel and Cremilleux [31] proposed a missing value imputation method, which extends the concept of Robust Association Rules Algorithm (RAR) for the databases consisting of multiple missing values. Chen et al. [32] applied selective Bayes classifier for the classification on incomplete data. Elshorbagy et al. [33] employed the principles of chaos theory to estimate the missing stream flow data. Nouvo [34] employed fuzzy c-means for data imputation.

Dempster et al. [35] designed the expectation maximization (EM) algorithm by using correlated data variables to estimate the missing samples in a multivariate data. Figueroa et al. [3] proposed a method to impute missing observations in the multivariate data using a genetic algorithm (GA) that minimizes an error function, derived from their covariance matrix and vector of the means. Recently, Ankaiah and Ravi [1] proposed a hybrid two stage imputation method to impute missing data. In the first stage, K-means algorithm is used to replace the missing values with the obtained cluster centers. In the second stage, MLP was used to refine the resultant approximate values. MLP was trained as a regression model by taking one incomplete variable as the target variable and the remaining variables as inputs. This procedure was repeated for the number of variables having the missing values.

This paper presents the power of the local learning capability by using in a clustering algorithm. An important objective of the paper is to perform sensitivity analysis on the influence of *Dthr* parameter on the clustering results.

III. Evolving Clustering Method

ECM is a one-pass, fast clustering method based on normalized Euclidean distances. It can be applied in two modes: on-line and off-line mode. The on-line method was employed in [36, 37] for time-series prediction. The off-line ECM is an extension of on-line ECM i.e. ECM with constrained optimization. We applied on-line ECM for our experiment to resolve the problem of missing values.

A. On-line Evolving Clustering Method

Without any optimization, ECM is a fast, one-pass algorithm for a dynamic estimation of the number of clusters in a set of data and for finding their current centers in the input data space [36]. It is a distance based clustering method. In any cluster, the maximum distance between an example point and the corresponding cluster center, is less than a threshold value, *Dthr*, a user-defined parameter. It affects the number of clusters to be estimated [36]. During clustering, if radius of any cluster exceeds the threshold value (*Dthr*), then we will stop the expansion of that particular cluster. We will see in the further section that how *Dthr* plays a major role in getting better accuracy. The procedure of ECM [36] is described as follows:

Step 0: Create the first cluster center C_1 by simply taking the position of the first example from the input data stream as the first cluster center Cc_1, and setting a value 0 for its cluster radius Ru_1.

Step 1: If all examples of the data stream have been processed, the algorithm is finished. Else, the current example x_i, is taken and the distances* D_{ij}, between this example and all the n already created cluster centers Cc_j,

$D_{ij} = ||x_i - Cc_j||$, j= 1 to n, are calculated.

Step 2: If there is a cluster center (centers) Cc_j, for j= 1 to n, so that the distance value, $D_{ij} = ||x_i - Cc_j||$ is equal to, or less than, the radius Ru_j, it is assumed that the current example x_i belongs to a cluster C_m with the minimum of these distances:

$$D_{im} = ||x_i - Cc_m|| = \min (D_{ij}),$$

Where: $D_{ij} <= Ru_j$, j= 1 to n.

In this case, neither a new cluster is created, nor any existing cluster is updated and the algorithm returns to Step 1, else it goes to next step.

Step 3: Find a cluster C_a (with a center Cc_a and a cluster radius Ru_a) from all n existing cluster centers through calculating the values $S_{ij} = D_{ij} + Ru_j$, j=1 to n, and then select the cluster center Cc_a with the minimum value S_{ia}:

$S_{ia} = D_{ia} + Ru_a = \min \{ S_{ij} \}$, j=1 to n.

Step 4: If S_{ia} is greater than 2 * *Dthr*, the example x_i does not belong to any existing clusters. A new cluster is created in the same way as described in Step 0, and the algorithm returns to Step 1.

Step 5: If S_{ia} is not greater than 2 * *Dthr*, the cluster C_a is updated by moving its center, Cc_a, and increasing the value of its radius, Ru_a. The updated radius Ru_a^{new} is set to be equal to $S_{ia}/2$ and the new center Cc_a^{new} is located on the line connecting the new input vector x_i and the cluster center Cc_a, so that the distance from the new center Cc_a^{new} to the point x_i is equal to Ru_a^{new}. The algorithm returns to Step 1.

* In this paper, the distance, between vectors x and y, is calculated as a normalized Euclidean distance, defined as follows:

$$\|x - y\| = \sqrt{\sum_{i=1}^{q} (x_i - y_i)^2} \; / \sqrt{q}$$

IV. ECM Based Imputation

ECM is one of the simplest unsupervised learning algorithms, which aids to solve missing data problem by its substantial local learning capability. The procedure of imputation is as follows (See fig. in last page i.e. page no. 7) :

- Divide a dataset in two parts: sets of complete and incomplete records.
- Perform ECM with the set of complete records and identify all the cluster centers.
- Attribute value, say x_k, in an incomplete record is imputed by the corresponding value of the attribute in the center of the nearest cluster by measuring the Euclidean distance between the incomplete record excluding the missing value and the cluster centers excluding the value in the same position. The Euclidean distance is measured by using the following formula:

$$D_j = \sum_{i=1; i \neq k}^{n} \left| x_i - c_j \right|^2$$

where, j – Number of cluster centers.

n – Number of complete components in each record.

To measure the effectiveness of the imputation, compute the mean absolute percentage error (MAPE) [38] for incomplete records as follows.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{xi - p_i}{xi} \right|$$

Where, n is the number of missing values in incomplete records, p_i is the predicted value and x_i is the actual value.

V. Experimental Setup

The effectiveness of the ECM-Imputation method for data imputation is tested on several regression, classification and banking datasets. Some of the datasets are the standard datasets from UCI Machine learning repository. Initially these datasets have no missing values, so we conducted the experiments by deleting some values from the original datasets randomly. We created missing values in all variables in a dataset except in a class or target variable.

Every dataset is divided into 10 equal folds and 9 folds undergo clustering and the tenth one is left out which had missing values. Since the dataset considered here do not have missing values, in order to create missing values, we randomly deleted nearly 10% of the values (cells) and ensured that at least one cell from every record is deleted. The same procedure is applied for all the remaining folds. Thus, in 10 fold cross validation, we have 10 folds with different missing records. For each fold the complete records are separated from the total records and used for the clustering. We applied the ECM algorithm on the complete dataset and missing value of the attribute of incomplete record was imputed by the corresponding value of the attribute of the nearest cluster center. For all datasets, we compared the average MAPE values (over 10 folds) of the proposed method with a hybrid method [1], which is based on K-means and MLP.

VI. Dataset Description

In this paper, we analyzed several regression, classification and banking datasets. Where, the regression datasets are Boston housing, forest fires, auto mpg and body fat.

The Boston housing dataset describes the housing values in the suburbs of Boston and it contains 506 records, 13 attributes. The Boston housing dataset was obtained from [39]. The forest fires dataset was obtained from [40].The Auto MPG dataset was obtained from [41]. The Auto MPG dataset concerns city-cycle fuel consumption in miles per gallon, to be produced in terms of multi valued discrete and continuous attributes. The body fat dataset was obtained from [42]. It lists the estimates of percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men.

The classification datasets are wine, Pima Indians, iris and Spectf. The wine dataset [43] contains 13 attributes and 178 records. Pima Indians dataset [44] was taken from National Institute of Diabetes and Digestive and Kidney Diseases in the year 1990. The Iris dataset [45] contains three classes of 50 instances each, where each class refers to a type of iris plant. The Spectf dataset [46] contains data on cardiac Single Proton Emission Computed Tomography (SPECT) images. Here each patient classified into normal and abnormal categories. The Spectf dataset contains 267 SPECT image sets (patients) which were processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature patterns were created for each patient.

The banking bankruptcy datasets are Spanish, Turkish and UK bankruptcy apart from a UK credit dataset. The Spanish bank dataset is obtained from [47]. Spanish banks dataset contains the list of banks which were bankrupt and non-bankrupt, so the target variable class contains bankrupt and non-bankrupt classes. This dataset contains 66 banks where 37 went bankrupt and 29 healthy banks. Turkish banks dataset is a bankruptcy prediction dataset where it contains patterns of several banks in which some are bankrupt and some other are non-bankrupt. It was obtained from [48], which is available at [49]. .This dataset contains 40 banks where 22 banks went bankrupt and 18 banks are healthy. The UK bankruptcy dataset is taken from [50]. This dataset contains 60 patterns among which 30 are healthy and 30 bankrupt. Each pattern corresponds to each bank. UK credit dataset consists of 1225 patterns of the customers applied for credit product [51]. In 1225 patterns 323 are of customers with very less credit i.e., bad customers and 902 are of good customers.

VII. Results And Discussion

We applied the ECM-Imputation method on all the datasets. The effectiveness of the proposed imputation method is tested on 4 regression, 4 classification and 4 banking dataset. We conducted experiment on a large range of *Dthr* values from 0.001 to 0.999, in steps of 0.001, in order to see which value of *Dthr* would perform better imputation compared to K-Means+MLP.

A. *Comparision of our results from K-Means+MLP and sensitivity analysis of Dthr value for imputation*

We compared our results with that of K-Means+MLP [1] and Mean imputation presented in Table I. Table I indicates that ECM-Imputation outperformed K-Means+MLP in 11 out of 12 datasets and outperformed Mean imputation in 10 out of 12 datasets in terms of average MAPE value. As can be observed from Fig. 1 through 12, ECM imputation is executed for a large range of 999 *Dthr* values and picked out that *Dthr* value, which performed best imputation. ECM imputation is not able to perform best for larger *Dthr* values. At first, MAPE value oscillated up and down due to increment of *Dthr* value but after a certain value, MAPE remained constant for the rest of the *Dthr* values. Since, number of clusters and value of cluster centers was not altering after certain range (We will discuss the range of *Dthr* values in next section), MAPE value did not change after a certain range. This sensitivity analysis of the influence of *Dthr* is performed on 12 datasets. Later, we resorted to Wilcoxon signed rank test in order to test the statistical significance of these results.

Wilcoxon two tailed signed rank test [52] is also performed at 1% level of significance to test the statistical significance of the results. The last column of Table 1 presents the results of Wilcoxon signed rank test. The critical value from the table [53] for N=10 is 3 at 1% level of significance. According to the Wilcoxon signed rank test, if the computed value is less than or equal to the critical value, then it is statistically significant.

Accordingly, one remarkable observation is that the ECM algorithm alone outperformed the K-Means+MLP hybrid algorithm presented in [1]. Further, we found that Mean imputation achieved better accuracy than the ECM based method in 2 datasets viz., Prima Indian and UK bankruptcy. But K-Means+MLP yielded better result in only one dataset. In Table I, bold digits denote that the average MAPE obtained by the proposed method is better than both Mean and K-Means+MLP imputation.

B. *Limiting the range of Dthr value*

Since, the number of clusters in ECM depends on the *Dthr* value and wide ranges of Dthr values are possible. In our experiment, we varied *Dthr* value from 0 to 0.108, as other values of Dthr yielded bad results. We kept the *Dthr* value as constant for all folds of the dataset and opted those results, which are the best among all the *Dthr* values. We depicted in the influence of *Dthr* on the MAPE values in Fig.1 through 12.

It can be observed from Fig. 1 through 12 that MAPE value remained constant after 0.430. Therefore, it is recommended that *Dthr* value should not exceed 0.430.

Table I. Mape Values For Datasets

	Mean	**K-Means +MLP**	**ECM Imputation**		**Wilcoxon signed rank test on ECM vs K-Means+ MLP**
			MAPE	*Dthr*	
Auto mpg	59.7	23.75	**18.03**	0.057	2.16
Body fat	11.61	7.83	**6.31**	0.081	1.86
Boston Housing	37.77	21.01	**17.84**	0.055	1.65
Forest fires	24.728	26.61	**22.29**	0.016	2.36
Iris	23.57	9.41	**5.27**	0.001	2.36
Prima Indian	24.022	29.7	27.16	0.047	0.94
Spanish	55.53	39.91	**31.98**	0.024	0.94
Spectf	14.85	12.14	**10.21**	0.082	2.16
Turkish	66.007	33.01	**27.90**	0.061	0.73
UK bankruptcy	37.07	30.96	46.14	0.087	2.47
UK Credit	28.43	32.17	**27.40**	0.001	1.04
Wine	29.99	21.58	**15.61**	0.108	2.57

Even though, MAPE value is constant after 0.430 but maximum *Dthr* value for better accuracy in our experiment is 0.108. That ECM outperformed the hybrid should not be surprising owing to similar experience in [54].

VIII. Conclusion And Future Work

We observed through our experiments that special local learning based imputation through ECM is efficient when compared to the hybrid of local learning and global approximation based imputation. We tested ECM-Imputation on 12 datasets in the framework of 10 fold cross validation. Several experiment had been conducted on a large range of *Dthr* values from 0.001 to 0.999, in steps of 0.001, in order to see which value of *Dthr* would perform better imputation compared to K-Means+MLP. By these experiments, we reached on the conclusion that *Dthr* values play vital role in ECM imputation. Therefore, it is essential to select an appropriate *Dthr* value; however, selection of an optimal *Dthr* value is an exhaustive search process. An exhaustive search of *Dthr* value increases its time complexity, so, any optimization algorithm will have required for selection of appropriate *Dthr* value. We also performed the Wilcoxon signed rank test to test the statistical significance of the computed values. Based on that, we conclude that the proposed approach can be used as viable alternative for the data imputation. Future work includes hybridizing ECM with other machine learning techniques in order to obtain more accurate imputations and employing some evolutionary optimization techniques to obtain optimal *Dthr* value.

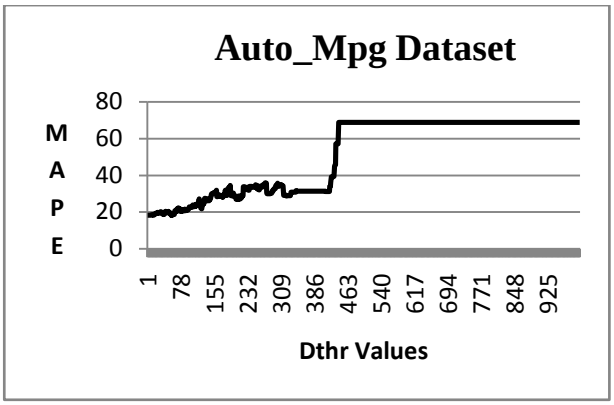

Fig. 1. Influence of *Dthr* value on MAPE results

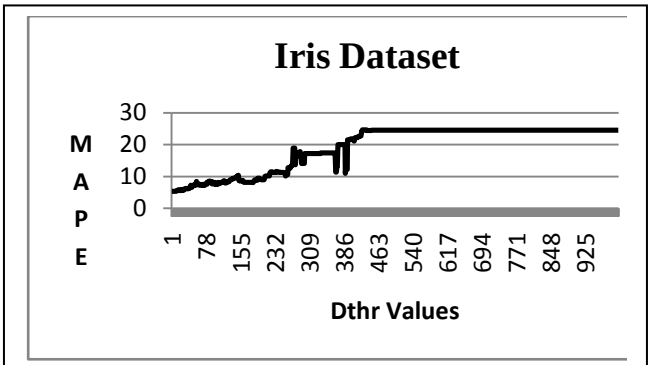

Fig. 5. Influence of *Dthr* value on MAPE results

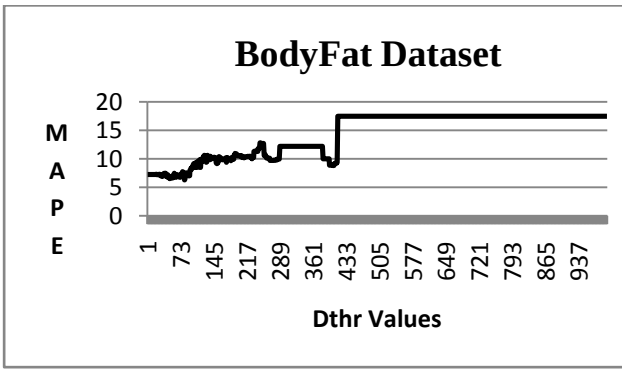

Fig. 2. Influence of *Dthr* value on MAPE results

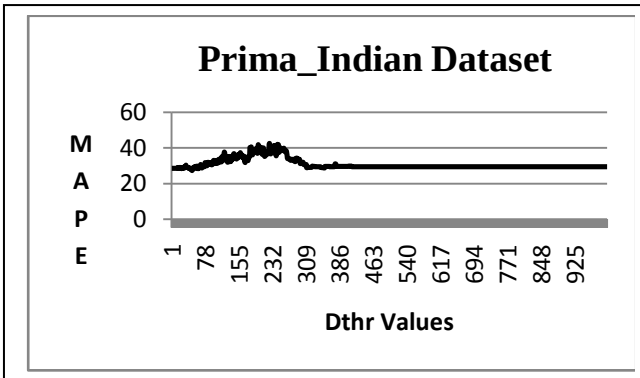

Fig. 6. Influence of *Dthr* value on MAPE results

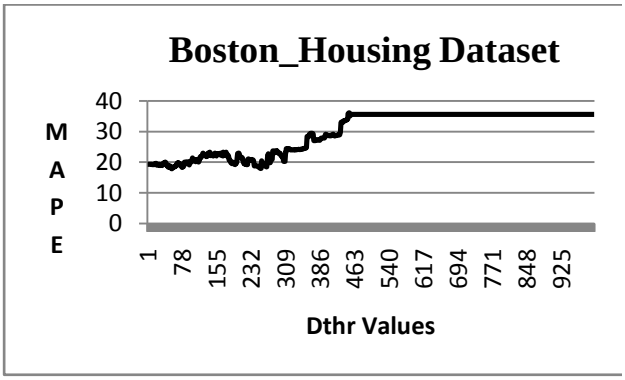

Fig. 3. Influence of *Dthr* value on MAPE results

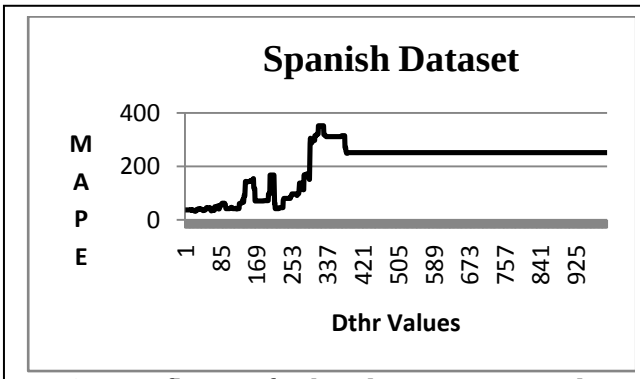

Fig. 7. Influence of *Dthr* value on MAPE results

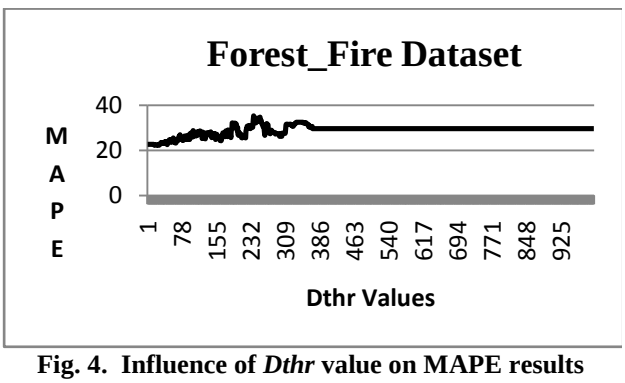

Fig. 4. Influence of *Dthr* value on MAPE results

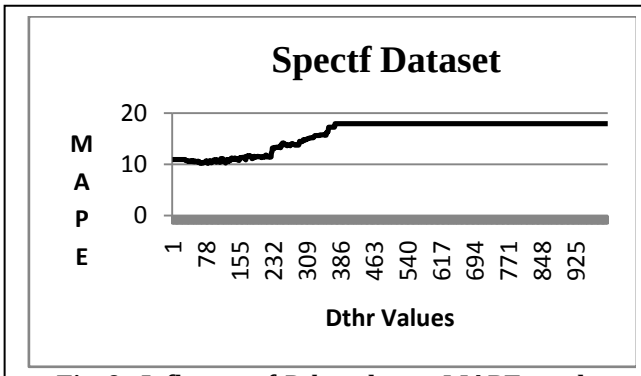

Fig. 8. Influence of *Dthr* value on MAPE results

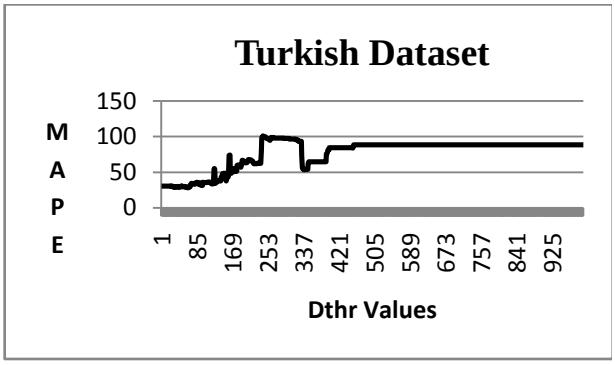

Fig. 9. Influence of *Dthr* value on MAPE results

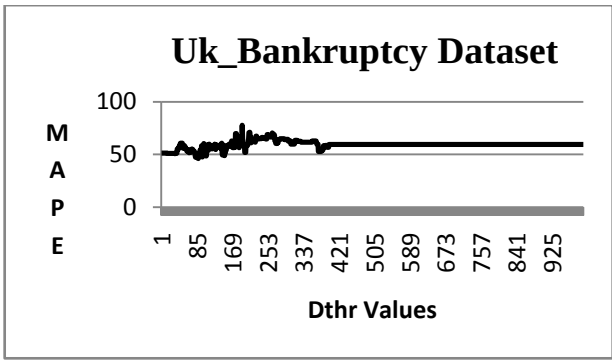

Fig. 10. Influence of *Dthr* value on MAPE results

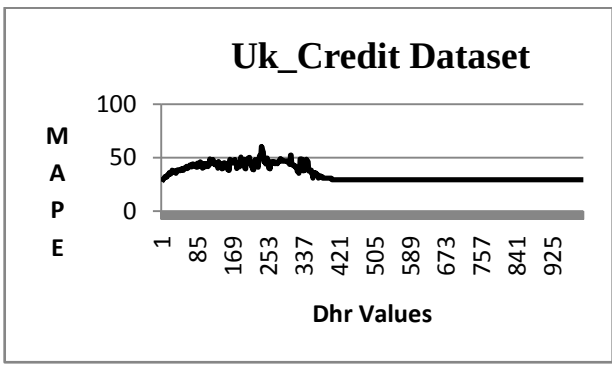

Fig. 21. Influence of *Dthr* value on MAPE results

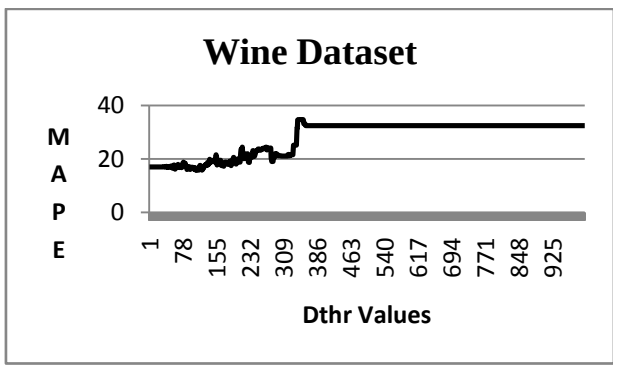

Fig. 32. Influence of *Dthr* value on MAPE results

REFERENCES

[1] N. Ankaiah and V. Ravi, "A novel soft computing hybrid for data imputation", In Proceedings of the 7th international conference on data mining (DMIN), Las Vegas, USA, 2011.

[2] M. Abdella and T. Marwala, "The use of genetic algorithms and neural networks to approximate missing data in database", In IEEE 3rd international conference on computational cybernetics, vol. 3, pp. 207-212, 2005.

[3] J. C. F. García, D. Kalenatic and C. A. L. Bello, "Missing data imputation in multivariate data by evolutionary algorithms", computers in Human Behavior, vol. 27, pp. 1468-1474, 2011.

[4] K.. J. Nishanth, V. Ravi, N. Ankaiah and I. Bose, "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts", Expert Systems with Applications, vol. 39(12), pp. 10583-10589, 2012.

[5] R. J. A. Little,. & D. B. Rubin, Statistical Analysis with Missing Data, New York: Wiley, 1987.

[6] R. J. A. Little,. & D. B. Rubin, Statistical analysis with missing data (2nd ed.).Hoboken, NJ, USA: Wiley-Inter science, 2002.

[7] I. G. Sande, Hot-deck imputation procedures. Incomplete Data in Sample Surveys, New York: Academic Press, 3, 339-349, 1983.

[8] A. Dempster, N. Laird, & D. Rubin, Maximum likelihood from incomplete data via the he Royal Statistical Society, Series B,39(1), 1–38, 1977.

[9] D. B. Rubin, Multiple imputation for nonresponse in surveys. Wiley, New York, 1987.

[10] R. B. Kline, "Principles and Practice of Structural Equation Modeling", Guliford Press, New York, 1988.

[11] Q. Song, M. Shepperd, "A new imputation method for small software project data sets", Journal of Systems and Software, vol. 80(1), pp. 51–62, 2007.

[12] J. L. Schafer, "Analysis of incomplete multivariate data", Florida, USA: Chapman & Hall, 1997.

[13] W. S. DeSarbo, V. R. Rao, "A constrained unfolding methodology for product positioning", Marketing Science, vol. 5(1), pp. 1-19, 1986.

[14] G. Batista and M. C. Monard, "A study of K-nearest neighbor as an imputation method", In Second International Comference on Hybrid intelligent systems, Santiago, Chile, Soft Computing System: Design Management and Application, IOS press, pp. 251–260, 2002.

[15] P. Merlin, A. Sorjamaa, B. Maillet and A. Lendasse, "X-SOM and L-SOM: A double classification approach for missing value imputation", Neurocomputing, vol. 73, pp. 1103–1108, 2010.

[16] B. Gabrys, "Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems", International Journal of Approximate Reasoning, vol. 30, pp. 149–179, 2002.

[17] A. Gupta and M. S. Lam, "Estimating missing values using neural networks", Journal of the Operational Research Society, vol. 47(2), pp. 229–238, 1996.

[18] J. Jerez, I. Molina, J. Subirates and L. Franco, "Missing data imputation in breast cancer prognosis", In Proceedings of the 24th IASTED international conference on biomedical engineering (BioMed'06), Anaheim, CA, USA, 2006.

[19] G. Batista and M. C. Monard, "Experimental comparison of K-nearest neighbor and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data", Technical Report, University of Sao Paulo, 2003.

[20] H. Liu and S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining", Journal of Systems and Software, vol. 85 (5), pp. 1067-1074, 2012.

[21] T. Samad and S. A. Harp, "Self-organization with partial data network", Computation in Neural Systems, vol. 3, pp. 205–212, 1992.

[22] P. C. Austin and M. D. Escobar, "Bayesian modeling of missing data in clinical research", Computational Statics & Data Analysis, vol. 49 (3), pp. 821-836, 2005.

[23] S. Nordbotten, “Neural network imputation applied to the Norwegian 1990 population census data”, Journal of Official Statistics, vol. 12, pp. 385–401, 1996.

[24] P. K. Sharpe and R. J. Solly, “Dealing with missing values in neural network based diagnostic systems”, Neural Computing & Applications, vol. 3(2), pp. 73–77, 1995.

[25] S. Y. Yoon and S. Y. Lee, “Training algorithm with incomplete data for feed-forward neural networks”, Neural Processing Letters, vol. 10, pp. 171–179, 1999.

[26] A. Gupta and M. S. Lam, “Estimating missing values using neural networks”, Journal of the Operational Research Society, vol. 47(2), pp. 229–238, 1996.

[27] T. R. Nkuna and J. O. Odiyo, “Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks”, Physics and Chemistry of the Earth, Parts A/B/C, vol. 36(14-15), pp. 830-835, 2011.

[28] E. L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello and M. D. Cubiles-de-la-Vega, “Missing value imputation on missing completely at random data using multilayer perceptrons”, Neural Networks, vol. 24 (1), pp. 121-129, 2011.

[29] M. Marseguerra and A. Zoia, “The auto-associative neural network in signal analysis II. Application to on-line monitoring of a simulated BWR component”, Annals of Nuclear Energy, vol. 32(11), pp. 1207–1223, 2002.

[30] T. Marwala and S. Chakraverty, “Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm”, Current Science India, vol. 90(4), pp. 542–548, 2006.

[31] A. Ragel and B. Cremilleux, “MVC—a preprocessing method to deal with missing values”, Knowledge Based Systems, vol. 12, pp. 285-291, 1999.

[32] J. Chen, H. Huang, F. Tian and S. Tian, “A selective Bayes Classifier for classifying incomplete data based on gain ratio”, Knowledge Based Systems, vol. 21 (7), pp. 530-534, 2008.

[33] A. Elshorbagy, S. P. Simonovic and U. S. Panu, “Estimation of missing stream flow data using the principles of chaos theory”, Journal of Hydrology, vol. 255 (1), pp. 123-133, 2002.

[34] A. G. Di Nuovo, “Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario”, Expert Systems With Applications, vol. 38 (6), pp. 6793-6797, 2011.

[35] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm”, Journal of Royal Statistical Society, vol. 39(1), pp. 1–38, 1977.

[36] Q. Song and N. Kasasbov, “Dynamic Evolving Neural-Fuzzy Inference System (DENFIS): On-line Learning and Application for Time-series Prediction”, Proc. 6th International Conference on Soft Computing, 696 – 701, Iizuka, Fukuoka, Japan, October, 2000.

[37] Q. Song and N. Kasasbov, “ECM — A Novel On-line, Evolving Clustering Method and Its Applications”, Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems, Berlin, 2001, pp. 87-92.

[38] B. E. Flores, “A pragmatic view of accuracy measurement in forecasting”, Omega, vol. 14(2), pp. 93-98, 1986.

[39] http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data

[40] http://archive.ics.uci.eduml/machine-learning-databases/forest-fires/forestfires.csv

[41] http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data

[42] http://lib.stat.cmu.edu/datasets/bodyfat

[43] http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

[44] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabete

[45] http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

[46] http://archive.ics.uci.edu/ml/machine-learning-databases/spect

[47] http://www.tbb.org.tr/english/bulten/yillik/2000/ratios.xls

[48] I. Olmeda , E. Fernandez, “Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction”, Computational Economics, vol. 10, pp. 317-335, 1997.

[49] S. Canbas, A.Caubak, S. B. Kilic, “Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case”, European Journal of Operational Research, vol. 166, pp. 528–546, 2005.

[50] M. J. Beynon, M. J. Peel, “Variable precision rough set theory and data discretisation:an application to corporate failure prediction”, Omega, vol. 29, pp. 561–576, 2001.

[51] L. C. Thomas, D. B. Edelman, J.N. Crook, “Credit scoring and its applications”, SIAM, Philadelphia, 2002.

[52] F. Wilcoxon, “Individual comparisons by ranking methods”, vol. 1, pp. 80-83, 1945.

[53] www.sussex.ac.uk/Users/grahamh/RM1web/WilcoxonTable2005.pdf .

[54] K. Huang, Learning From Data Locally and Globally, “A Ph.D thesis submitted in Department of Computer Science & Engineering”, The Chinese University of Hong Kong, 2004.