Improving Accurate Candidates for Missing Data Using Benefit Performance of (ML-SOM)

Abeer Abdullah AL-Mohdar * **Mohamed Abdullah Bamatraf****

Abstract

Missing data is one of the major challenges in extracting and analyzing knowledge from dataset. The performance of training quality was affected by the appearance of missing data in a datasets. For this reason, there is a need for a quick and reliable method to find possible solutions in order to provide an accurate system. Therefore, the previous studies provided robust ability of Self Organizing Map (SOM) algorithm to deal with the missing values [6, 20]. However, it has a drawback such as an error rate(ERR) in the missing values that increase huge dataset. This study is mainly based on the projection of unsupervised Multilayer SOM (ML-SOM) for missing values. The global methodology presented the combination of advantages of the proposed ML-SOM to obtain a precise method with various missing rates: 5%, 10% and 20%. The experiments were conducted by adopting two types of commonly used data benchmarks (IRIS and Breast-Cancer) from Weka 3.9 machine learning tool. The new proposed method ML-SOM provides a minimum Error Rate (ERR) and high accuracy (ACC)in small and large datasets compared to other standard classifier types (Bayes-Net, K-means and SOM).

Keywords: Self-organizing-map, Multi-layer self-organizing-map, missing data, Data mining, Neural networks, Machine learning.

Introduction:

Missing data is a serious problem in real world applications, researchers face the challenge of dealing with such data. Missing data is a common problem that has become a rapidly growing area and it is the current focus of this research[26].

Most of the collected datasets from real-life domains contain missing data which deemed to be very significant in affecting and extracting knowledge from these datasets. Data must be treated before working, at the pre-processing phase. The preprocessing stage assists to make the data more accurate, consistent and precise for a processing stage in order to make appropriate decisions during the building and evaluating of the statistical analysis and data mining models.

Numerous techniques are used to predict and fill them, due to its problem that can negatively influence the quality of the data analysis in additional to the accuracy of the generated models.

Consequently, there are many researchers focusing on developing appropriate techniques to overcome this challenge like: [18, 7].

Broadly speaking, the classification of missing values in datasets relies on three factors: attribute, instance, and missing values, which occur randomly in attributes and instances.

Methods of missing data:

1- Substitution methods: after classifying data according to a pre-defined criteria (also called hot-deck methods); individuals should use data from similar observations.

2- Prediction methods: which adopt statistical models standard on the database (mod imputation, mean imputation, regression approaches, etc.) rather than using manual correcting it [23].

Ignoring, deleting and Imputation are techniques for replacing missing data with substituted values. If an important feature is missed for a particular instance, it can be estimated from the data that are present by using these imputations[26]. In this respect there are several methods in machine learning that are employed to deal with imputation missing value like : k-means, Support Vector Machines(SVM) , Artificial Neural Network (ANN), and SOM [16, 24, and 8]. The Self-Organizing Map(SOM) was proposed by Professor Teuvo Kohonen (1982). It was a neural network trained with unlabeled data (unsupervised learning).

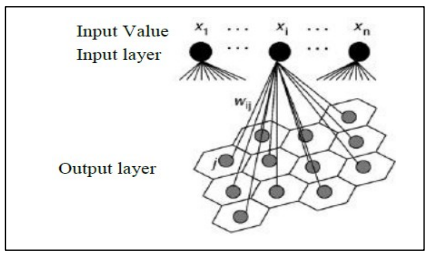
To represent a high-dimensional data, all data were mapped into one point, namely node (winner node) in the map and the distances of the items in the map reflect similarities between the items in the

* Information Technology Department, Faculty of Computers and Information Technology, Hadhramout University.

** Information Technology Department, Faculty of Computers and Information Technology, Hadhramout University. Received on 4/2/2020 and Accepted for Publication on17/6/2020

map and those topological compressing information but it preserves the relationships of the original data[1].

The generic structure is displayed in the following figure (1) and it allows for operation on nearly any kind of input data as long as it is metric on those provided areas. Moreover, their ability to learn in an unsupervised fashion enables them to adapt to even completely unforeseen input patterns without any supervisory intervention. These advantages make the SOM popular choices for tasks of missing value imputation [12, 5, 19], mutation[23], image compression[4], image color quantization[3] etc.



Figure(1): Single layer SOM

Classification of missing values cases:

A large number of contributing authors adopted the property of SOM to discover and manipulate the problem of missing data that are based on techniques of machine learning algorithm such as [25] that used properties of SOM and tested it on different artificial problems. MLP was able to represent and classify structured patterns [9, 17] with better performance. This becomes possible especially when networks are trained through knowledge by unsupervised learning like [11, 2, 10, 13, and 21].

In this paper, the techniques and properties of SOM algorithm are proposed as a special model to overcome the problem of missing data by adopting the special technical ML-SOM in a big dataset. An experiment was conducted on how ML-SOM were trained by using knowledge obtained by the information in SOM algorithm [15].

This paper begins describing the principle of the SOM [14, 22] then it gives the special SOM structure in ML-SOM. The main motivations behind the works presented are to train the process and to generate SOM knowledge. Furthermore, the researchers selected new disciplines to demonstrate the applicability of SOM to adapt with the nature of missing data imputation domain in a totally different and independent way.

SOM algorithm:

The basic SOM comprises of M neurons which is located usually on a 2-D grid that is either hexagonal or rectangular. Each neuron is competitive to each other to be a winning node. For this reason SOM is also known as a competitive network rather than learning task where has (d-dimensional) feature vector $W_{ij} = [W_{i1}, \ldots, W_{id}]$. In this case, it is better to choose random values for the initial weight vectors W_j, and assign a small positive value to the learning rate parameter α.

Structure of Multilayered SOM(ML-SOM):

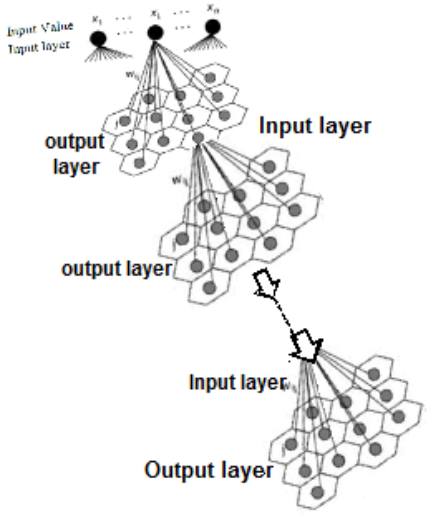
The ML-SOM consists of multiple layers, each layer comprised of SOM model . All number of units in each layer are described at successive levels, resulting in a pyramidal structure. The number of representative vectors are generated in each layer that is proportional to the number of neural units in the output layer for each SOM. The input layer receives an input from the external layer and propagates the input to all neural units in the next competitive layer.

The output from any given layer is converted into the input for the next layer as shown in figure(2) that represents our model. The same process is repeated until the top layer is reached. Thus, there are a fewer numbers of neural units in a layer at a higher level. Therefore, each weight vector represents a larger cluster. Hence, the representation produced at a higher level in the ML-SOM corresponds to a higher level of abstraction of the input data making the ML-SOM well suited for hierarchical range data representation.

The researchers are training this model by four different standard classifiers(**Bayes-Net, K-means, SOM and ML-SOM**), which are selected in order to cover the techniques broadly applied in the representative statistical strategies as in the accuracy including the two types of big and small data benchmarks **(**IRIS**,** Breast-Cancer).

The iterative (ML-SOM) training algorithm can be stated as follows:

Step 1: Set iteration t = 0 for L no of SOMs .



Figure(2): ML-SOM Model

Step 2: Randomly select a sample data vector Xi and compute Euclidean distances between Xi and all feature to find Best Matching Unit (BMU) at iteration p(for each SOM) using the norm of minimum distance usual measure in "Equation(1)":

$$E = \min_j \left\| X - W_j(p) \right\| = \sqrt{\sum_{i=1}^{n} \left[X_i - W_{ij}(p) \right]^2} \quad \ldots(1)$$

$$j = 1, 2, \cdots$$

The letter (n) is the number of neurons in the input layer, while (m) is the number of neurons in the SOM layer.

Step 3: Update the weight of BMU neuron and its neighbor neurons to move its feature vector towards the data vector in "Equation(2)":

$$W_{ij}(p+1) = W(p) + \Theta(P)\alpha(P)\left(X(p) - W_{ij}(p)\right) \quad \ldots(2)$$

where Θ is restraint due to distance from BMU and it is usually called the neighborhood function, $\alpha(t)$ is the learning rat, $W_{ij}(p)$is the weight repairing in pth iteration.

Step 4: Return to step 2 until the feature map stops changing, or no noticeable changes occur in the feature map and when L=t.

After processing SOMs layers, the result should be a spatial organization of the input data organized into similar regions.

Evaluation metrics ML-SOM:

To evaluate the performance of this proposed module by calculating the Accuracy (Acc) which is used to measures a classification of dataset, as a ratio between the correct predictions of a classified samples from a total number of samples as shown in "Equation 3":

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots(3)$$

The complement of the accuracy metric which is used to be incorrect predictions or misclassification rate is the Error rate (ERR) .

This represents the number of misclassified samples from both positive and negative samples, and it is calculated as "Equation 4":

$$ERR = (FP+FN)/(TP+TN+FP+FN) \quad \ldots(4)$$

These letters P and N are the numbers of positive and negative samples respectively in "Equation 3, 4".

Additionally, the researchers consider sensitivity and specificity as two kinds of accuracy, where the TP is deemed as actual positive samples whereas the FN is for actual negative samples. Sensitivity depends on TP and FN which are in the same column of the confusion matrix. Similarly, the specificity metric depends on TN and FP which are in the same column; hence, both sensitivity and specificity can be employed for evaluating the classification performance with imbalanced data [8].

Experimental results:

This module was trained by adding missing data into all pattern that is selected randomly in small datasets(IRIS) and large datasets (Breast-Cancer) in various percentages of missing value (5%, 10% and 20%) .

This takes place with the aim of estimating the performance in the proposed module ML-SOM with other standard classifiers (**Bayes-Net**, K-means and SOM)

which give high ACC than other classifiers especially in 20% missing values. Then ERR is calculated for each datasets, which are ratio minimum error of misclassified data in ML-SOM.

When the stopping criteria is satisfied, there no weight update between the input pattern and the target value that take minimum ERR experiment possible variability of relative performances of classifiers across datasets.

All of the experiment results are summarized below in table (1) of small datasets (IRIS), and large datasets (Breast-Cancer) is summarized in table (2).

Table (1): classifiers of small datasets with missing values.

	Classifier's	Dataset with Missing Value	ACC	ERR
Small data set (IRIS)	**Bayes-Net**	5%	0.89	0.11
		10%	0.82	0.18
		20%	0.63	0.37
	K-means	5%	0.93	0.07
		10%	0.79	0.21
		20%	0.66	0.34
	SOM	5%	0.94	0.06
		10%	0.86	0.14
		20%	0.82	0.18
	ML-SOM	5%	0.96	0.05
		10%	0.88	0.12
		20%	0.90	0.1

Table (2): classifiers of Big datasets with missing values.

	Classifier's	Dataset with Missing Value	ACC	ERR
Big data set (Breast-Cancer)	**Bayes-Net**	5%	0.91	0.09
		10%	0.89	0.11
		20%	0.73	0.27
	K-means	5%	0.95	0.05
		10%	0.85	0.15
		20%	0.88	0.12
	SOM	5%	0.95	0.05
		10%	0.83	0.17
		20%	0.90	0.1
	ML-SOM	5%	0.91	0.09
		10%	0.86	0.14
		20%	0.93	0.07

Conclusion:

The presented results showed that the appearance of missing data in a large dataset has a negative effect of performance in a training quality. The proposed module enhances the traditional SOM algorithm, in order to make the most of its ability to deal with missing data and increase its accuracy and reducing the error rate as little as possible. Also when the size of the dataset repository increases, it is mainly a good classification feature of the SOM, and it is dependent on this new model. ML-SOM could simulate the activation functions of each layer then modified these layers when updating the corresponding weight that links the incomplete dataset(data with missing value) in an input layer in order to reach the target outputs.

To conclude, the estimation of the proposed module ML-SOM could give high accuracy of a correct predication of samples. Also, the least expected error rate in the presented uncorrected predication samples with missing values in different proportion(5%, 10% and 20%) gave an excellent results compared to other standard classifiers pertaining to both statistical and machine learning which is shown in table (1 and 2).

References:

1- Bassani, H. and Araujo, A., 2019. A neural network architecture for learning word–referent associations in multiple contexts, Elsevier Ltd., 117, 249-267.

2- Bengio, Y., 2009. Learning deep architectures for AI. Foundations and trends in Machine Learning, 2(1), 1–127.

3- Celton, M., Malpertuy, A., Lelandais, G., and Brevern, A., 2010. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics 11, 1-16.

4- Chen, C., Grennan, K., Badner, J., Zhang, 2008. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics 9, 1-12.

5- Cottrell, M. and Letr´emy, P., 2005. Missing values: Processing with the kohonen algorithm. Applied Stochastic Models and Data Analysis, Brest,489–496.

6- Cottrell, M. and Letremy, P., 2005. Missing values: Processing with the kohonen algorithm. Applied Stochastic Models and Data Analysis, Brest, France, 489–496.

7- Fessant, F. and Midenet, S. , 2002. Self-Organising Map for Data Imputation and Correction in Surveys, Neural Computing & Applications. Neural Comput & Applic, 10, 300–310.

8- Garcia, V., Mollineda R.A. and Sanchez, J.S., 2010. Theoretical analysis of a performance measure for imbalanced data. 20th International Conference on Pattern Recognition (ICPR), IEEE, 617–620.

9- Hammer, B., 2002. Recurrent networks for structured data-a unifying approach and its properties. Cognitive Syst. 3(2), 145–165.

10- Hinton, G. and Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

11- Hinton, G., 2007. Learning multiple layers of representation. Trends in cognitive sciences, 11(10), 428-434.

12- Kang, H. and Yusof, F., 2012. Application of Self-Organizing Map (SOM) in Missing Daily Rainfall Data in Malaysia. International Journal of Computer Applications (0975 – 888), 48(5).

13- Kmnimura, R., 2011. Self-enhancement learning: target-creating learning and its application to self-organizing maps. Biological cybernetics, 1- 34.

14- Kohonen, T., 1998. Self Organization and associative memory springer Series in Information Sciences. 2nd (Berlin : Springer).

15- Kohonen, T., 2010. The Self Organizing Map. Information Sciences Springer Verlag, New York .30, 312.

16- Little, R.J., 2011. Calibrated Bayes, for Statistics in General, and Missing Data in Particular. Statistical Science, 26(2),162-174.

17- Little, R.J.A.and Rubin, D.B., 2002. Statistical Analysis with Missing Data, Wiley Interscience, 2.

18- Marshall, A., Altman, DG., Royston, P., and Holder, RL., 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. MC Med Res Methodol, 10(7).

19- Nkiaka, E., Nawaz, N.R. and Lovett, J.C., 2016. Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment. Lake Chad basin. Environmental Monitoring and Assessment, 016-5385-1.

20- S. Antti, Maillet, B., Merlin, P., and Lendasse, A., 2007. Som+eof for finding missing values. European Symposium on Artificial Neural Networks, 115–120.

21- Saitoh, F. , 2016. An ensemble model of self-organizing maps for imputation of missing values. 2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA), 172.

22- Samad, T. and Harp S., 1992. Self Organisation with partial data. Network. 3, 205–212.

23- Schmitt, P., Mandel, J. and Guedj, M., 2015. A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics&Biostatistics,DOI:10.472/21556180.100 022,6-1.

24- Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. Australasian Joint Conference on Artificial Intelligence, Springer, 1015–1021.

25- Sperduti, A. and Starita, A., 1997. Supervised neural networks for the classification of structures. IEEE Trans. Neural Networks, 8(3) , 714–735.

26- Tahani, A. and Sasi, S., 2016. Proper imputation techniques for missing values in data sets. International Conference on Data Science and Engineering (ICDSE), 1-5.

تحسين دقة التصنيف للبيانات المفقودة بالاستفادة من فعالية أداء خوارزمية التنظيم الذاتي متعدد الطبقات

عبير عبدالله المحضار محمد عبدالله بامطرف

الملخص

تعد البيانات المفقودة إحدى أهم التحديات الرئيسية في استخراج وتحليل المعرفة في قواعد البيانات الكبيرة، وقد أثر ظهور البيانات المفقودة في قواعد البيانات الكبيرة في أداء جودة التدريب، ولهذا السبب هناك حاجة إلى طريقة سريعة وموثوقة وإيجاد حلول من أجل توفير نظامٍ دقيقٍ. لذلك ،ومن خلال الدراسات السابقة لخوارزمية خريطة التنظيم الذاتي (SOM) التي تميزت بقدرتها للتعامل مع القيم المفقودة [6 ، 20]. ورغم ذلك، فقد لوحظ أن خوارزمية التنظيم الذاتي لها عيب يكمن في أن معدل الخطأ(ERR) عند التعامل مع القيم المفقودة يزداد كلما ازداد حجم البيانات مما يفقدها التميز في إعادة بناء الأنظمة الدقيقة. اعتمدت هذه الدراسة بشكل أساسي على بروز أثر SOM متعدد الطبقات غير الخاضع للرقابة –ML) (SOM بالتعامل مع القيم المفقودة بكفاية عالية. فقد جمعت المنهجية العامة للبحث التي تم تقديمها مزيجاً بين مزايا ML–SOM المقترحة بالحصول على دقة عاليةِ (ACC) للخوارزمية في ظل وجود معدلات مختلفة للقيم المفقودة بين: 5٪ ، 10٪ و 20٪. وطبقت الخوارزمية على نوعين من قواعد البيانات المعيارية شائعة الاستخدام (قزحية العين وسرطان الثدي) صغيرة الحجم وكبيرة الحجم من أداة التعلم الآلي Weka 3.9. وقد أظهرت الخوارزمية الجديدة المقترحة ML–SOM الحد الأدنى من معدل الخطأ ERR ودقة عالية ACC في كلتي قواعد البيانات صغيرة وكبيرة الحجم بمقارنتها مع غيرها من مصنفات الخوارزميات القياسية المعتمدة.(Bayes-Net and K-means ,SOM)

الكلمات الدالة: خريطة التنظيم الذاتي ، خريطة التنظيم الذاتي متعدد الطبقات ، البيانات المفقودة ، الشبكات العصبية ، التعلم الآلي.