EJONS International Journal on Mathematic, Engineering and Natural Sciences ISSN 2602 - 4136

Article Arrival Date	Article Type	Article Published Date
30.10.2020	Research Article	15.12.2020

Doi Number: http://dx.doi.org/10.38063/ejons.365

CLUSTERING PRACTICES IN MISSING VALUE DATA SETS

Serpil SEVİMLİ DENİZ

Van Yüzüncü Yıl University, Gevaş Vocational School, Computer Programming Department.
ORCID: 0000- 0002-8559-1107

H. Eray ÇELİK

Van Yüzüncü Yıl University, Faculty of Economics and Administrative Sciences, Department of Econometrics.

Çağdaş Hakan ALADAĞ

Hacettepe University, Faculty of Science. Department of Statistics.

ABSTRACT

Missing data is when one or more values cannot be obtained in the data sets. The purpose of cluster analysis is to provide summary information to the researcher by classifying the data according to their similarities and to reduce the number of data that is too much to less. In this study, the performances of the three clustering methods are compared using different missing data rates in eleven separate data sets consisting of numerical and nominal data. The correct clustering rates of the data were examined by decreasing the data at five percent, ten percent, fifteen percent, twenty percent, twenty five percent and thirty percent of the data sets completely and randomly. The methods whose working performance were tested using missing data are k-means, one of partitioned clustering methods and self-organizing maps, one of artificial neural network-based clustering methods - Self Organization Map (SOM) and linear vector segmentation model - Learning Vector Quantization (LVQ). According to the results of the analysis; it is observed that as the missing data rate increases, the correct cluster rate decreases. It was observed that the LVQ method performed better in four data sets with two sets of nominal and numerical data, while the SOM method performed better clustering in the other seven data sets consisting of numerical data.

Keywords: Missing Data, Clustering, SOM, LVQ, k-means.

INTRODUCTION

For accurate and reliable analysis, it is desirable that all data are complete, but in all systems that can be measured and monitored, the presence of more or less missing data is inevitable. There may be many reasons for data loss. Missing data sets pose problems in many statistical analyzes. Removing missing-valued observations from the data set causes the sample volume to decrease and the statistical power of the analyzes performed decreases (Akpınar, 2014). Cluster analysis can also be considered as a data modeling problem, as it focuses on finding relationships and meaningful shapes between objects in the data set.

k-means

The k-means method uses k prototypes, the centers of the clusters, to define data. They are determined by minimizing the error sum of squares. k-means (Macqueen, 1967) is one of the most used methods for classical but partitioned clustering. It groups the data set in k groups. Grouping is done by minimizing the sum of squares of distances between data points and the respective cluster centers. The logic of the method depends on the iterations of these steps; First, determining the coordinate of the center is based on the evaluation of the distance of each object from the centers and the final grouping of objects based on the minimum distance (Macqueen, 1967).

The need for the user to know the number of clusters in advance can be seen as a disadvantage. Moreover, when data sets are different in size, density and non-spherical shapes the k-means method cannot produce the desired good result. The main disadvantage is the sensitivity to outliers and noise (Han et al., 2012).

Artificial neural networks used in clustering

1) Self Organization Map (SOM)

Self-organized map is one of the most popular neural network models. It is in the category of competitive learning networks. Self-organized map (SOM) is an unsupervised learning model. This means that little information about the properties of the input data is sufficient. SOM can be used for data clustering without knowing the class membership of the input data. The concept of self-organized map was developed by Teuvo Kohonen (Kohonen, 1993). It provides a topology protection mapping from high-dimensional space to map units. Map units or neurons often form a two-dimensional frame. Mapping is a mapping from high-dimensional area to a plane. Points close to each other in the entry area are mapped to the close map units in SOM. SOM can therefore be used as a cluster analysis tool in high-dimensional data. Also, SOM has the ability to generalize. The generalization feature means that the network can recognize or characterize inputs that it has never encountered before. A new entry is represented by map unit (Kohonen, 2001).

SOM Learning Algorithm

The dependent variable is not used in the education of the network. As the input vectors in the data set are entered into the network, the network is self-organizing and reference vectors are created. This algorithm is as follows (Zontul et al., 2004).

Symbols used in this algorithm:

$x_n = x_{n1}, x_{n2}, ..., x_{nm}$: Input vectors for data metris x obtained from $m*n$ consisting of m feature and n record (metris x obtained from $m*n$

$w_j = w_{1j}, w_{2j}, ..., w_{mj}$: Reference vectors for output neurons j for m kernel weight

$d(i, j)$: Square of Euclidean distance to output neuron in (i, j) coordinate of input vector

J : output neurons with the input vector closest.

α : learning coefficient

h : neighbouring function

c : winning neuron

Algorithm (Zontulve et al., 2004; Vatansever, 2008).

Step 0: Assign initial values to w_{ij} coefficients

Determine the topological neighborhood (R) parameters

Determine learning coefficient (α) parameters

Step 1: For input vector and weight vector, calculate Euclidean distances in the following way;

$$d(w_j, x_n) = \sqrt{\sum_i (w_{ij} - x_{ni})^2}$$

Step 2: Find the j winning neuron where $d(w_j, x_n)$ is the minimum for all neurons.

Step 3: Find the neighborhood J of the winning neuron j for the neighborhood parameter R.

Step 4: Update reference vectors as follows for all output neurons (J) in the specified neighborhood of j in iteration t.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)h_{ci}(t)(x_{ni} - w_{ij}(t))$$

Step 5: Update the learning coefficient.

Step 6: Decrease the topological neighborhood parameter at specified times (Ultsch & Siemon, 1990; Larose, 2005; Van Hulle, 2012).2)

2) Linear Vector Segmentation Model (Learning Vector Quantization-LVQ)

The LVQ network was developed in 1984 by Tuevo Kohonen. The basis of LVQ networks is the Kohonen layer in the SOM model developed by Kohonen. The basic philosophy of the LVQ model for the classification problem is to map an n-dimensional vector to a series of vectors. It is one of the frequently preferred networks due to its fast results and high performance (Öztemel, 2003). LVQ networks consist of three layers. The first layer, the input layer, does not process information. The incoming information forms the input layer. The second layer is the intermediate layer, also called the Kohonen layer. In this layer, the closest weight vector to the input set is determined as in SOM. Each element in this layer represents a reference vector. The input vector is mapped to reference vectors formed by weights between the input layer and the Kohonen layer. In the third layer, the output layer, the class to which the input belongs, is determined (Baş, 2006).

Structure of the LVQ Network

In the LVQ networks, each neuron in the input layer is associated with all neurons in the Kohonen layer. Neurons in the Kohonen layer are associated with a single neuron in the output layer. The weights between the Kohonen layer and the output layer are equal to 1, fixed and these weights do not change. The education of the network is carried out only by changing the weights between the Kohonen layer and the input layer, that is, the values of the reference vectors. Thanks to these changes, reference vectors are determined to classify the entries into the correct classes. When training the network, it is determined whether the output produced by the network is just correct instead of the value at each iteration. Only values close to the vector (winning vector) input vector (weights of the network for this vector) are changed (Kröse et al., 1996). The outputs of each neuron element in both the Kohonen layer and the output layer take binary values, and only one neuron element has 1 output value, and the others are 0. The output value of 1 indicates that the entries belong to the class entered. The quantization algorithm is used in the training of LVQ networks. The purpose of training a network is to find the reference vector closest to the input vector in each iteration, that is, to set the center of the cluster and to minimize the quantitative errors of all the input vectors in the training set. Reference vectors are weight values that, as previously noted, connect neurons in the Kohonen layer to neurons in the input layer. During learning, only the weight values of the reference vectors are changed. This is done using the Kohonen learning rule. The Kohonen learning rule is based on the principle that the neuron elements in the Kohonen layer compete with each other. The competitive criterion is calculated by the Euclidean distance between the input vector and the weight vectors (reference vectors). The neuron closest to the input vector wins the competition. There are two conditions for the winning neuron (Cozart, 1996). In the first case, the winning neuron is a member of the correct class.

In this case, the respective weights are brought closer to the input vector. This is done for the same neuron to win again when the same sample is shown to the network again. In this case, the weights are changed. The learning coefficient is reduced to 0 over time in a monotonous way. This is because the input vector stops when it is very close to the reference vector and does not retract again. Failure to do so will result in further reversal. In the latter case, the winning neuron is the wrong class. In this case, the weight vector must be subtracted from the input vector. The purpose of this is that the same neuron element does not win the same sample when it comes next time. Weights are then changed. The reduction of the learning coefficient over time is also valid here. The weights between the Kohonen layer and the output layer do not change during training. The outputs of the neuron elements in the Kohonen layer are multiplied by the weight values that connect these neuron elements to the output layer to calculate the output of the network. When the outputs of the network are determined, it is questioned whether the output is classified correctly. The answer is to change the weights that connect the neuron to the input layer in the Kohonen layer. Therefore, LVQ networks are in a reinforced learning class. These procedures are continued until all the samples in the training set are correctly classified. When all are correctly classified, learning takes place (Elmas, 2010).

DATA SETS

Eleven data sets were used from the UCI data repository. These are given in Table 1

Table 1. Data sets used in the study

Data Set	Size	Cluster Number	Data Type
Iris	5*150	3	Numeric
Breast Cancer	10*699	2	Numeric and Nominal
Blood Transfusion	5*748	2	Numeric and Nominal
Diabets	7*768	2	Numeric and Nominal
Wine	15*178	3	Numeric
Abalone	9*4177	3	Numeric
Credit Card	21*1000	2	Nümeric and Nominal
Ionosphere	35*351	2	Numeric
Censor	13*2212	3	Numeric
Segment	20*1500	7	Numeric
Gas	128*13980	6	Numeric

METHOD

In this study, it is aimed to compare the performances of the methods when applied in data sets with missing values by using k-means from partitioned clustering techniques and SOM and LVQ methods from artificial neural network based clustering techniques. In order to make a qualitative comparison, the real data sets that are widely used in making such comparisons are used in the literature. In the first stage, five small, five medium and one large data sets containing numerical and nominal data were selected, and the performance of clustering methods according to the correct classification rates were examined in the full data sets. In the second stage, five percent, ten percent, fifteen percent, twenty percent, twenty five percent and thirty percent data from these data sets were completely decreased at random and new data sets were created, and the performance of clustering methods according to the correct classification rates were examined. The results obtained in the first step in order to list the methods that perform the best clustering were evaluated according to the correct classification rates. In the second step, the correct classification rates were matched and the status of the clustering methods were evaluated using the paired-t test.

TESTS AND RESULTS

In data sets containing complete and missing data, cluster performances of k-means, SOM and LVQ clustering methods were evaluated at full, 5 percent, 10 percent, 15 percent, 20 percent, 25 percent and 30 percent completely random missing data rates. Data sets are classified as small, medium and large. In this study, cluster performances were tested in cases where the data were missing completely and randomly. The data sets used in the study were selected as multivariate, numerical or nominal data. Therefore, the scope of these results is limited to numerical and nominal data, whose data is completely and randomly missing. In this study, it was investigated how some of the clustering methods cluster in the presence of missing data. K-means, one of partitioned clustering methods and SOM and LVQ from ANN based clustering methods are handled. As a result of the study carried out, it is observed that there is not much change in the correct classification rates of the k-means method in all data sets used. Accordingly, it can be said that the k-means method can tolerate missing data.

Table 2. Data sets and methods are compared

Data Set	k-means	LVQ	SOM
Iris	81.67	80.87	**88.67**
Breast Cancer	76.03	**86.15**	66.01
Blood Transfusion	61.21	**65.17**	53.86
Diabets	62.67	**66.78**	57.40
Wine	77.60	73.29	**94.15**
Abalone	47.04	46.34	48.07
Credit Card	56.60	**58.60**	51.50
Ionosphere	65.33	62.88	**70.31**
Censor	68.26	60.52	**68.72**
Segment	56.53	15.73	**66.93**

Since the gas data set is very large, the methods could not be compared with each other, but SOM, k-means and LVQ ranking were determined as the best method ranking according to the correct classification rates.

DISCUSSION

Contrary to the generally used data assignment methods, studies that argue workability in data sets with missing data started to develop after 2013. In a study of Orczyk and Porwik (2013), they show the dangers of filling in missing data - especially medical data. Juhola and Laurikkala (2013), in their study to determine the effect of missing values on true positive ratios and classification accuracy in five data sets, and where they used KNN (k-nearest neighbours), Discriminant Analysis and Naif Bayes methods, despite up to 20-30% missing values in the two-class data sets, they showed that better results can be produced as much as there are not any missing values. When the correct classification rates obtained in the study conducted are evaluated, it was observed that as the missing data rates increased, the correct classification rates decreased but there were no significant differences. Zhu and Shi (2018) proposed the full use of observed data to reduce the error caused by filling missing values instead of data assignment methods in their study, where they proposed a new support vector machine algorithm for missing data. They used accuracy, F score, and Kappa statistics to verify method. In the study carried out, instead of filling the missing data with different clustering methods, the working performance of the methods with different missing data rates were tested using the correct classification rates.

CONCLUSION

In eleven different data sets used according to the results of the analysis, it is observed that the correct classification rate decreases as the rate of missing data generally increases. The best clustering order varies depending on the content of the data sets. It is seen that the data which contains nominal data from the analyzed data sets and the output consists of two sets are best clustered with LVQ. It forces data to be 1 or 0 with supportive learning after the Kohonen layer due to its LVQ structure. This confirms the results of the data sets examined. In data sets containing numerical data, it is seen that SOM method shows the best clustering performance. When the big data set is analyzed, it is seen that the increase of the missing data rate does not affect the correct clustering rates much.

REFFERENCES

Akpınar H. (2014). Data. Papatya Puplishing, Istanbul.

Baş, N. (2006). Artificial Neural Networks Approach and an Application (Unpublished Master Thesis). Mimar Sinan University Institute of Science, Istanbul.

Cozart M. T. (1996). Evaluation of ‘The Neural Gas’ Network Vector Quantization and Approximation Components (Master of Science Thesis). The University of Tennessee, Knoxville.

Elmas, Ç. (2010). Artificial Intelligence Applications. Seçkin Publishing, Ankara.

Han, J., Kamber, M., Pei, J. (2012). Data Mining Concepts and Techniques. Morgan Kaufman.

Juhola, M., ve Laurikkala, J. (2013). Missing Values: How Many Can They Be To Preserve Classification Reliability? *Artificial Intelligence Review*, 40(3): 231– 245.

Kröse, B. ve Smagt, P.V.D. (1996). An Introduction to Neural Networks. The University of Amsterdam.

Kohonen, T. (1993). Things you haven’t heard about the Self-Organizing Map in Neural Networks. IEEE International Conference, 1147–1156.

Kohonen, T. (2001). Self-Organizing Maps, Springer Series in Information Sciences. Springer Berlin.

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. 5th Berkeley Symp. Mathemtaical Statistics&Probability, 1: 281– 297.

Orczyk, T. ve Porwik, P. (2013). Influence of missing data imputation method on the classification accuracy of the medical data. https://yadda.icm.edu.pl/baztech/ element. Access Date: 03.04.2018.

Ultsch A. ve Siemon H.P. (1990.) Kohonen's self organizing feature maps for exploratory data analysis. InProc. INNC'90, Int. Neural Network Conf., 305-308, Dordrecht, Netherlands. Kluwer.

Van Hulle M. M. (2012). Self-Organizing Maps. Springer, Berlin.

Vatansever M. (2008). Use and Application of Visual Data Mining Techniques in Cluster Analysis (Master's Thesis) Yıldız Technical University Institute of Science and Technology, Istanbul.

Zhu ve Shi, (2018). A Novel Support Vector Machine Algorithm for Missing Data. https://dl.acm.org/doi/10.1145/3194206.3194214. Access Date: 05.07.2019

Zontul, M., Kaynar, O. ve Bircan, H. (2004). Using SOM Neural Networks for Turkey's Import from a Study on the Clustering of the Country. Cumhuriyet University, Journal of Economics and Administrative Sciences, 5 (2): 47- 68.