IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Clustering Incomplete Mixed Datasets by using Extended Squeezer Algorithm and Finding Incomplete Set Mixed Dissimilarity (ISMD)

M.V. Jagannatha Reddy[1] D. Ramachandra Reddy[2], M. Mahesh Kumar[3]

[1]Dept. of Computer Science & Engineering, Aditya College of Engineering, Madanapalle
[2]Dept. of Computer Science & Engineering, Aditya College of Engineering, Madanapalle
[3]Dept. of Computer Science & Engineering, Aditya College of Engineering, Madanapalle

Available online at: www.ijcseonline.org

Accepted: 27/Sept/2018, Published: 30/Sept/2018

Abstract: Clustering mixed datasets is one of the challenging task. Traditional algorithms like k-prototype algorithm is used for mixed dataset, but is limited to only complete datasets. In any dataset missing values are common. To handle such missing values or incomplete mixed datasets we use extended squeezer algorithm, which includes the new dissimilarity measure ISMD that is incomplete set mixed dissimilarity for numerical and categorical attribute values. In this method we consider dissimilarities in the missing values and in this extended squeezer algorithm it not only cluster the incomplete dataset, it also need not to input the missing values and need not to initialize any clusters at the beginning. This method is compared with traditional k-prototype algorithm on benchmark datasets. The experimental results shows that the ISMD using extended squeezer algorithm gives better accuracy than the traditional k-prototype algorithm and also it overcomes the limitation of initial clusters. This method is implemented by using Python programming. The results shows that there is significant improvement in the clustering results.

Key Words: Incomplete set mixed dissimilarity, k-prototype, extended squeezer algorithm, Python programming

I. INTRODUCTION

Clustering is a process of dividing the datasets into groups called clusters. The objects within the clusters are more similar and the objects between the clusters are more dissimilar. Clustering is applied in almost all fields for example to analyze their business trends, for customer partition, in pattern recognition, medical field and so on. In the real world most of the datasets are of incomplete sets. Handling such an incomplete dataset is a challenging task. Some of the techniques presently using is the missing value sets are removing from the list. As a result of removing missing data accuracy may affect. It is very important that to deal with incomplete mixed numerical and categorical datasets, otherwise accuracy of clustering decreases.

The remainder of this paper organized as, in section 2 is discussed about the related work done by the previous researchers, in section3 discussed materials and methods, section 4 describes performance evaluation and section 5 gives conclusion.

II. RELATED WORK

In our previous paper we presented a clustering algorithm for mixed numerical and categorical dataset using similarity weight and filter method [1]. Now we extend our work to find clustering algorithm for incomplete mixed numerical and categorical dataset. For this let us study some of the works done by previous researchers. K-means algorithm is one well known earliest clustering algorithm for numeric data. In this it uses mean as its center of the cluster. Since it clusters only numerical datasets, Huang [2] proposes a k-modes algorithm which instead of using means, it uses modes as its center based on the frequency of the attribute values. So that this can be used for categorical attribute values to extend k-means [3] to use simple matching distance measure. However even k-modes algorithm is not able to handle mixed datasets. Hence Huang [4] gave a k-prototype algorithm which includes the properties of both k-means algorithm and k-modes algorithm. Even in the k-prototype algorithm have some disadvantages that have to initialize clusters, dissimilarity measure and isolated points. Choosing k elements as initial clusters affect the results and it increases the number of iterations and it takes more time to complete the process [5]. So we need to overcome this problem. This problem can be overcome by using extended squeezer algorithm. Also this algorithm improves the dissimilarity measure and isolated points [6], employed the object closest to the center point of mean.

Two types of clustering categories are listed below
a) Traditional Clustering algorithms

1) Partitioning：K-means, PAM, CLARA, K-modes, EM, CLARANS, ISODATA.
2) Hierarchical：BIRCH, CURE, Chameleon.

3) Grid：STING, Wave cluster, CLIQUE.
4) Density：DBSCAN, OPTICS, DENCLUE.
5) Model：COBWEB, CLASSIT, LVQ, SOM.

b) New Clustering algorithms
1) Ownership relation：Granular, Uncertainty, Spherical Shell, Entropy.
2) Data preprocessing：Kernel, Concept.
3) Similarity：Spectral, Hybrid data, Dual Distance.
4) Update strategy：Data Stream, PSO.
5) High dimensional：Projection Pursuit, Subspace.
6) Integration with other science：Clustering Ensemble, Random Walk.

To handle incomplete data with missing values is always a difficult task. One of the method to handle missing values is deleting the entire row of the missing value but these methods produce biased results. So imputation of missing values is a challenging task.

Imputation is a technique through which imputes known or suitable value into the missing data using some algorithms. The principle of imputation adopts three methods, imputation based on statistics, roughest theory and data mining i.e classification and clustering. Chatzis Sortirios [7] uses c-means substitution method in which mean value is replaced by mean value. The second method, roughest theory utilizes the tolerance relation between the objects which can handle only categorical attribute values. the third method uses some data mining techniques classification and clustering. Wang Fengmei [8] used k-nearest neighbors imputation calculates the Euclidean distance between each object with missing values and other complete objects, and then chooses the k –objects with smallest distance to the incomplete object to calculate missing values with weighted average value. Xu Fang [9] proposed a modified shuffled frog leaping clustering method which based on k-means technique, this divides the data into complete data and incomplete data, then it assigns the imcomplete objects to the closest complete objects and fills the missing values with cluster centers. Takashi Furukawa, Shin-ichi Ohnishi, and Takahiro Yamanoi [10] used fuzzy c-means algorithm for mixed incomplete data using partial distance and imputation they showed that this can be applied on real dataset. Vaishali H. Umathe, Prof. Gauri Chaudhary [11] performed review on incomplete data and clustering, the imputation approach is used to fill the incomplete data sets and IFCMwUNC clustering algorithm, is used for clustering.

Many methods of imputation have still limitations. Some of these imputation methods handles only numerical data or categorical data and some other methods simply fills the missing values by mean values [12]. Because of these many errors may cause during the process. Also non missing values of incomplete objects are ignored in these imputation methods. They concentrate only on complete data.

To overcome the limitations of the above said algorithms we proposed extended squeezer algorithm uses new dissimilarity measure for incomplete objects with mixed numerical and categorical dataset. Also this algorithm overcomes limitation of initial k clusters initialization. After the clustering process the missing values can also be imputed based on the results. The implementations results shows that the extended squeezer algorithm gives better accuracy. And also overcomes the limitation of initial cluster center initialization.

Previously this method is implemented using R programming[15]. Now this method is implemented by using python programming.

III. MATERIALS AND METHODS

3.1 DATA SET

For this implementation we used thyroid dataset from UCI machine learning repository to evaluate the performance of the extended squeezer algorithm based on the dissimilarity measure for incomplete dataset with mixed numerical and categorical dataset. The thyroid dataset is collected by new south wales institute of 3428 tuples each of is described by 15 categorical and 6 numerical attributes.

3.2 PROBLEM DESCRIPTION

Incomplete dataset S={U,A,V,f}, where U={$x_1,x_2,\ldots,x_n$}, A=$C \sqcup N=\{ak|k=1,2,\cdots,m\} \sqcup \{al|l=m+1, m+2, .., m+q\}$. The number of objects is *n*, the number of attributes is $m+q$, *C* is the data set of categorical attribute, *N* is the data set of numeric attributes, *V* is the set of all values. *f* represents the function $U \times A \rightarrow V$. In this paper, we set the missing values by "*".

3.3 INCOMPLETE SET MIXED DISSIMILARITY (ISMD)

Wu Sen, Chen Hong and Feng Xiaodong [13], Given the incomplete system $S=\{U,A,V,f\}$, *X* is a subset of *U*; xi and xj are two objects in X and |*X*| is the number of objects in *X*. Here incomplete set mixed dissimilarity (*ISMD*) of *X* combines (xi,) and (xi,) with weight is defined as:

$$ISMD_{(x_i,x_j)} = \begin{cases} \dfrac{W_c ISMDC(x_i,x_j) + w_n x ISMDN(x_i,x_j)}{w_c + w_n} \end{cases} \text{--(1)}$$

Where w_c represent the weight of the categorical attribute and w_n represent the weight of numerical attribute.

ISMDC(*xi,xj*) represent the degree of dissimilarity for categorical attribute and by rough set theory [14]. It is defined as

$$\delta_k(x_i,x_j)=\begin{cases}1 a_k(x_i)\neq a_k(x_j)\wedge a_k(x_i)\neq "*"\wedge a_k(x_j)\neq "*"\\ 0 a_k(x_i)=a_k(x_j)\vee a_k(x_i)="*"\vee a_k(x_j)="*"\end{cases} \quad --(2)$$

$$ISDC\left(x_i,x_j\right)=\frac{\sum_{k=1}^{m}\delta_k\left(x_i,x_j\right)}{m-\sum_{k=1}^{m}\delta_k\left(x_i,x_j\right)} \quad --(3)$$

Where (*xi*,) represents distance between *xi* and *xj* in categorical attribute *k*. consider m is the number of categorical attributes and q is the number of numerical attributes.

ISMDN(*xi,xj*) represents the dissimilarity degree of numeric attributes .Based on the distance of mini-max standardization and Minkowski distance, the formula is given:

$$d_l(x_i,x_j)=\begin{cases}\dfrac{|al(xi)-al(xj)|}{Max_l-Min_l} -> a_l(x_i)\neq "*"\wedge a_l(x_j)\neq "*"\\ 0 --> a_l(x_i)="*"\vee a_l(x_j)="*"\end{cases} \quad (4)$$

$$ISMDN(x_i,x_j)=\frac{\left(\sum_{l=m+1}^{m+q}d_l(x_1,x_2)^2\right)^{1/2}}{q-\left(\sum_{l=m+1}^{m+q}d_l(x_1,x_2)^2\right)^{1/2}} \quad (5)$$

Where *Maxl* represent the maximum and represent the minimum value of attribute *al* in *X*. and (*xi*,) is the standardized numeric distance and (*xi*,) is calculated through normalizing as above formula (5).

The incomplete set mixed dissimilarity (*ISMD*) can deal with incomplete data sets with mixed numerical and categorical attributes. Further, with no need for imputing mean values in advance, *ISMD* measures dissimilarity of data objects with missing values directly and these decreases the clustering errors.

3.4 EXTENDED SQUEEZER ALGORITHM FOR INCOMPLETE MIXED DATASETS

The main steps involved in clustering incomplete mixed numerical and categorical datasets using extended Squeezer algorithms is as follow,

Input: Incomplete dataset S = S={U,A,V,f} and minimum similarity s;
Output: k clusters and dataset s after imputation
while (D has unread tuple){
tuple = getCurrentTuple (D)
if (tuple.tid = = 1){
addNewClusterStructure
(tuple.tid) }
else{
for each existed cluster C
search the minimum incomplete set mixed dissimilarity between the object x_i and cluster by calculating the dissimilarity ISMD(x_i, c)
get the max value of similarity:
get the corresponding cluster index
if ISMD>=s
addTupleToCluster(tuple, index)
else
addNewClusterStructure
(tuple.tid) }
take all k clusters and impute the missing values by taking mean M_i for numeric values and Mode$_i$ for categorical values the formulas are as follow:

$$I_l=\left\{x_i\in x\middle| a_l(x_i)="*"\right\} --> (6)$$

$$M_i=\begin{cases}\dfrac{1}{|x|-|I_l|}\sum_{x_l\in I_l}a_l(x_i) --|x|>|I_l|\\ M_i=+\alpha --> |x|=|I_l| \quad (7)\end{cases}$$

Where $a_l(x_i)$ represents the value of object x_i on numeric attribute a_l

$$Mode_i=\begin{cases}\max(|a_k(x_i)|) --> a_k(x_i)\neq "*"\\ 0 --------> a_k(x_i)="*"\end{cases} \quad (8)$$

Where $a_k(x_i)$represents the value of object x_i on categorical data a_k

OutputClusteringResult ()

IV. PERFORMANCE EVALUATION

The performance of the algorithms and cluster accuracy is evaluated by using three metrics, RC, RMSSE and clustering accuracy r

(1) Accuracy rate in categorical attribute

For the categorical attribute, the accuracy rate for categorical attribute after imputation RC is the ratio of

average number of correct imputed categorical values per run and the starting number of missing values. it is given by

$$RC = C/(\gamma + m + n) \quad (9)$$

Where C is the number of correct imputed categorical values. RC indirectly tells the performance of clustering algorithm. Larger the RC the algorithm performance is more better.

(2) Root Mean Standard Squared error: RMSSE

Suppose for numeric attributes, the missing attribute values are v_1, v_2,…$v_{\gamma*q*n}$ and based on clustering algorithms imputed values are $V'1, V'2, \cdots, V'\gamma*q*n$. The root mean standard squared error RMSSE can be defined as,

$$RMSSE = \sqrt{\frac{\sum_{i=1}^{\gamma*q*n} (\square_i)^2}{\gamma * q * n}} \quad (10)$$

$$\Delta_i = \begin{cases} 1 ----> v_i' = "*" \\ \frac{v_i - v_i'}{Max - \min} \quad --> v_i' \neq "*" \wedge Max \neq Min \\ \Delta_i = 0 ---> v_i' \neq "*" \wedge Max = Min \end{cases} \quad (11)$$

where Max is the mean of maximum and Min is the minimum of attribute values with missing numeric values.

(3) Clustering Accuracy r

The quality of the clustering results is measured by using the formula,

$$r = \frac{\sum_{i=1}^{k} c_i}{n} \quad (12)$$

Where c_i is the number of objects in the correct clusters and k is the number of clusters.

V. EXPERIMENTAL RESULTS

During implementation the above algorithms are written and implemented in python programming environment. To evaluate the accuracy of the extended squeezer algorithm, we carried out several runs of traditional k-prototype and mextended squeezer algorithm with each missing rate of experimental datasets.

From the formula (9), the RC results of thyroid dataset is shown in the fig.1 below. It is observed the accuracy rate in categorical attribute RC of the extended squeezer algorithm is better than the initial one.

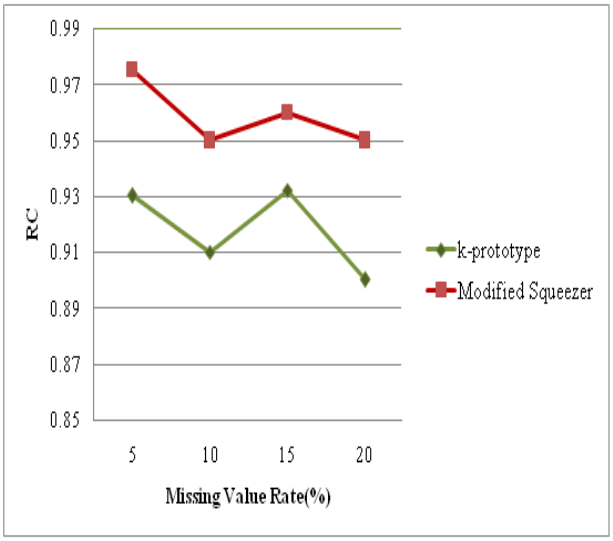

Figure.1 Comparison between k- prototype and extended squeezer algorithm in RC vs Missing value rate

From the figure1 it is seen that the accuracy rate in categorical attribute RC of the extended squeezer algorithm is higher than the traditional k-prototype with the increase in the missing value rate. The mean RC of the k-prototype is 0.970 compared to that of extended squeezer algorithm is 0.995. Thus the extended squeezer algorithm performed better than the k-prototype algorithm in clustering.

According to formula (10), figure 2 below shows the RMSSE results for the credit approval dataset. From that it is seen that the extended squeezer algorithm is better than the k-prototype algorithm.

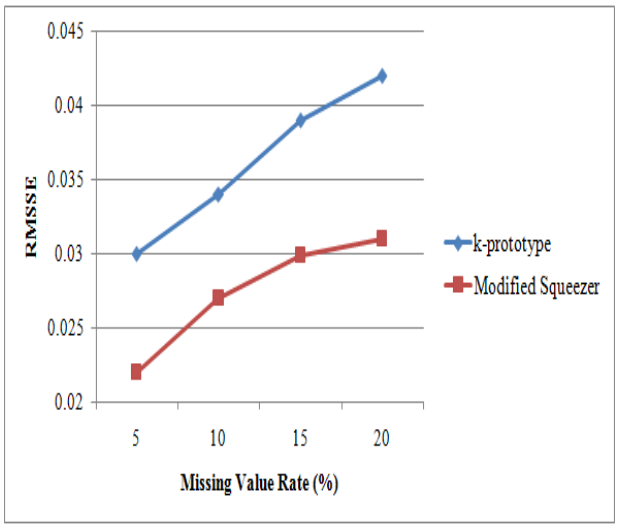

Figure 2:RMSSE comparison between k-prototype and extended squeezer

Finally the clustering accuracy r of k-prototype and extended squeezer algorithm on the credit approval dataset is shown in the figure below. It depends on the formula (12)

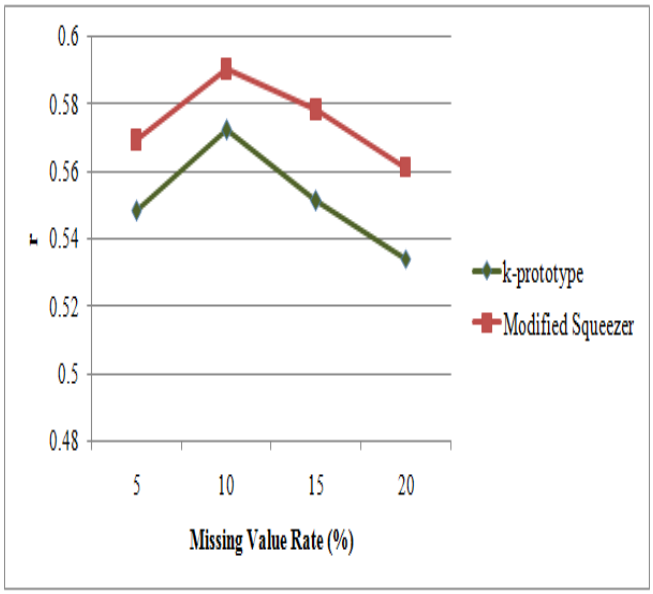

Figure 3: r comparison between k-prototype and extended squeezer

From the above figure 3 it is observed that extended squeezer algorithm is better than the k-prototype algorithms with the increase in missing value rate. As a conclusion we can say that for the high missing rate the extended squeezer algorithm performs better than the traditional k-prototype algorithm.

CONCLUSION

From the above experimental results implemented by using Python programming seen that the proposed ISMD and extended squeezer algorithm performed better in RC, RMSSE and r. The proposed method not only performs more accurately on incomplete data, but also it is effective in filling missing values. In this method it does not require filling incomplete data in advance, in this new method of clustering we impute the missing values after clustering results, so that the error rate in the clusters reduces. Also, in this algorithm it need not initialize the clusters at the beginning. Finally we conclude that extended squeezer algorithm performed better compared to k- prototype in terms of RC, RMSSE or r using python programming.

REFERENCES

[1] M.V.Jagannatha Reddy and Dr. B. Kavitha, "clustering mixed numerical and categorical dataset using similarity weight and filter method", International journal of Database Theory and Applications, vol-5, no-1 March- (2012), pp-121-134

[2] H. Zhexue, "Extension to the K-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, (1998), pp. 283-304.

[3] T. Covões and E. Hruschka, "A study of K-Means-based algorithms for constrained clustering", Intelligent Data Analysis, vol. 17, no. 3, (2013), pp. 485-505.

[4] H. Zhexue, "Clustering large data sets with mixed numeric and categorical values", Proceedings of the 1th pacific-Asia Conference on Knowledge Discovery & Data Mining. Singapore: World Scientific, (1997), pp. 21-34.

[5] W. Qian, W. Cheng and F. Zhenyuan, "Summary of k-means clustering algorithm", Electronic Design Engineering, vol. 20, no. 7, (2012), pp. 21-24.

[6] C. Dan and W. Zhenhua, "A K-prototypes Algorithm Based on Improved Initial Center Points", Computer Knowledge and Technology, (2010) November.

[7] C. Sotirios, "A fuzzy c-means-type algorithm for clustering of deal with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional", Expert Systems with Applications, vol. 38, no. 7, (2011), pp. 8684-8689.

[8] W. Fengmei and H. Lixia, "A Missing Data Imputation Method Based on Neighbor Rules", Computer Engineering, vol. 38, no. 21, (2012).

[9] X. Fang and Z. Guizhu, "Clustering algorithm based on Modified Shuffled Frog Leaping Algorithm and K-means", Computer Engineering and Applications, vol. 49, no. 1, (2013), pp. 176-180.

[10] Takashi Furukawa, Shin-ichi Ohnishi, and Takahiro Yamanoi "On a Fuzzy c-means Algorithm for Mixed Incomplete Data Using Partial Distance and Imputation" Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I, IMECS 2014, March 12 - 14, 2014, Hong Kong.

[11] Vaishali H. Umathe, Prof. Gauri Chaudhary. "A Review on Incomplete Data And Clustering" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, pp 1225-1227

[12] J. Twisk, M. de Boer, W. de Vente and M. Heymans, "Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis", Journal of Clinical Epidemiology, vol. 66, no. 9, (2013), pp. 1022-1028.

[13] Wu Sen, Chen Hong and Feng Xiaodong "Clustering algorithm for incomplete data sets with mixed numeric and categorical Attributes" IJDTA, vol. 6 No. 5 2013, pp 95-104.

[14] W. Guoyin, "Expansion in the theory of rough set in incomplete information system", Journal of computer research and development, vol. 33, no. 10, (2002), pp. 1239-1240.

[15] M..V.Jagannatha Reddy, Dr.B.Kavitha "Clustering Incomplete Mixed Numerical and Categorical Datasets using Modified Squeezer Algorithm International Journal of Computer Science and Engineering, E-ISSN:2347-2693, Vol-4, issue-5 pp-36-41 may-16

About Authors

Dr.M.V.Jagannatha Reddy completed B.E.,M.Tech.,Ph.D from reputed universities. He is having 21 years of teaching experience in various positions. Presently he is working as Associate Professor and Head of the CSE department in Aditya College of Engineering, Madanapalle. He presented and published 13 research papers in various international conferences. Also he published 13 research papers in various journals. His journals having 37 citations which includes 2 h-index and 1 i-index. He successfully completed UGC Minor research project worth Rs.2.05 lakhs.

D.Ramachandra Reddy completed M.Tech. from JNTUA. Presently working as Assistant professor in CSE department of Aditya College of Engineering, Madanapalle. He is having more than 8 years of teaching experience. His areas of interest are data Mining, Data structures.

M.Mahesh Kumar completed M.Tech. from JNTUH. Presently working as Assistant professor in CSE department of Aditya College of Engineering, Madanapalle. He is having more than 9 years of teaching experience. His areas of interest are Cloud Computing, Data Mining

