Optimal Number of Cluster Identification using Robust K-means for Sequences in Categorical Sequences

S.U. Patil[1], U.A. Nuli[2]

[1,2]*Computer Science and Engineering department, M. Tech, Textile and Engineering Institute, Ichalkaranji, Maharashtra, India*

Abstract - ***This paper presents a modified k-means algorithm for clustering. In traditional method of clustering number of clusters to be formed will be given at the start of the algorithm, which affects performance and efficiency of the algorithm. In Robust K-means for sequences optimal number of cluster will be predicted by removing noise cluster. Cluster validation, which is the process of evaluating the quality of clustering results, plays an important role for practical machine learning systems. Categorical sequences, such as biological sequences in computational biology, have become common in real-world applications. Different from previous studies, which mainly focused on attribute-value data, in this paper, we work on the cluster validation problem for categorical sequences. Clustering is defined as an unsupervised learning where the objects are grouped on the basis of some similarity inherent among them. The intension of this paper is to describe the clustering method which will give the optimal number of clusters in categorical sequences.***

***Key Words:* Clusterign , K-means, cluster validation index, categorical sequences, centroid.**

1. INTRODUCTION

Data mining is a process of deriving required data from a collection of large dataset and making analysis on collected data. The data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. In generic term this is known as classification. In classification when the classes of an object is given in advance is termed as supervised classification where as the other case when the class label is not tagged to an object in advance is termed as unsupervised classification. The unsupervised classification is commonly known as clustering. Clustering is important analysis techniques that is employed to large datasets and finds its application in the fields like search engines, recommendation systems, data mining, knowledge discovery, bioinformatics and documentation. Nowadays, the data being generated is not only huge in volume, but is also stored across various machines all around the world. The main purpose behind the study of classification is to develop a tool or an algorithm, which can be used to predict the class of an unknown object, which is not labeled.

Clustering problem cannot be solved by one specific algorithm but it requires various algorithms that differ significantly in their notion of what makes a cluster and how to efficiently find them. Generally clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Clustering is considered to be more difficult than supervised classification as there is no label attached to the patterns in clustering. The given label in the case of supervised classification becomes a clue to grouping data objects as a whole. Whereas in the case of clustering, it becomes difficult to decide, to which group a pattern will belong to, in the absence of a label.

The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values. This is often the case in many domains, where data is described by a set of descriptive attributes, many of which are neither numerical nor inherently ordered in any way. Moreover clustering categorical sequences is a challenging problem due to one more reason the difficulties in defining an inherently meaningful measure of similarity between sequences. Cluster validation, which is the process of evaluating the quality of clustering results, plays an important role for practical machine learning systems. Categorical sequences, such as biological sequences in computational biology, have become common in real-world applications. Without a measure of distance between data values, it is unclear how to define a quality measure for categorical clustering. To do this, we employ mutual information, a measure from information theory. A good clustering is one where the clusters are informative about the data objects they contain. Since data objects are expressed in terms of attribute values, we require that the clusters convey information about the attribute values of the objects in the cluster. The evaluation of sequences clustering is currently difficult due to the lack of an internal validation criterion defined with regard to the structural features hidden in sequences. To solve this problem, a novel cluster validity index (CVI) is proposed as a function of clustering, with the intra-cluster structural compactness and inter-cluster structural separation linearly combined to measure the quality of sequence clusters. Cluster validation, which is the process of evaluating the quality of clustering results, plays an important role in many issues of cluster analysis A partition-based algorithm for robust clustering of categorical sequences and the CVI are then assembled within the common model selection procedure to determine the number of clusters in categorical sequence sets. Currently, cluster validation remains an open problem due to the unsupervised nature of clustering tasks, where no external validation criterion is available to evaluate the result. Despite different aspects of cluster validation, we are interested in the problem of determining the optimal number of clusters in a data set, as most basic clustering algorithms assume that the number of clusters is a user-defined parameter, which, however, is difficult to set in practice.

2. RELATED WORK

K. A. Abdul Nazeer et al. [12] discuss in this paper about the one major drawback of the k-means algorithm. K-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Two phases includes in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean. An enhanced clustering method is discussed in this paper, in which both the phases of the original k-means algorithm are modified to improve the accuracy and efficiency. This algorithm combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. But still there is a limitation in this enhanced algorithm that is the value of k, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points.

G. Dong and J. Pei, [10] says algorithm is expected to produce high-quality results associated with the underlying cluster structures in the data set given *K*. This problem becomes difficult for categorical sequences, because, usually, they are infected with significant quantities of noise, which confuses the identification of cluster structures. The clusters containing less number of objects after applying a algorithms are considered as K-dependent noises. Such cluster-number-dependent noises would mislead the validation criterion to evaluate the clustering quality. The analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in each iteration. This repetitive process affects the efficiency of clustering algorithm.

R. Xu and D. Wunsch [11], in the existing partition-based methods, the performance of k-means algorithm which is evaluated with various databases such as Iris, Wine, Vowel, Ionosphere and Crude oil data Set and various distance metrics. proposed that due to the increment in the amount of data across the world, analysis of the data turns out to be very difficult task. To understand and learn the data, classify those data into remarkable collection. So, there is a need of data mining techniques. It is concluded that performance of k-means clustering is depend on the data base used as well as distance metrics. the comparative analysis of one partition clustering algorithm (k means) and one hierarchical clustering algorithm (agglomerative).On the basis of accuracy and running time the performance of k-means and hierarchical clustering algorithm is calculated using WEKA tools. This work results that accuracy of k-means is higher than the hierarchical clustering for iris dataset which have real attributes and accuracy of hierarchal clustering is higher than the k-means for diabetes dataset which have integer, real attribute. So for large datasets k means algorithm is good.

R. Amutha et al. [13] proposed that when two or more algorithms of same category of clustering technique is used then best results will be acquired. Two k-means algorithms: Parallel k/h-Means Clustering for Large Data Sets and A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets. Parallel k/h-Means algorithm is designed to deal with very large data sets. Novel K-Means Based Clustering provides the advantages of using both HC and K-Means. Using these two algorithms, space and similarity between the data sets present each nodes is extended, however, there is a general lack of adaptive mechanisms to deduce the reasonable number of clusters with the possible *K*-dependent noises identified, given that the user defined *K* likely deviates from the true number. Solution to this problem a novel cluster validation method which will provide robust clustering is used.

Navjot Kaur, Navneet Kaur[4]enhanced the traditional k-means by introducing Ranking method. Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means. The Ranking Method is a way to find the occurrence of similar data and to improve search effectiveness. The tool used to implemented the improved algorithm is Visual Studio 2008 using C#. The advantages of k-means are also analyzed in this paper. The author finds k-means as fast, robust and easy understandable algorithm. He also discuss that the clusters are non-hierarchical in nature and are not overlapping in nature. The process used in the algorithm takes student marks as data set and then initial centroid is selected. Euclidean distance is then calculated from centroid for each data object. Then the threshold value is set for each data set. Ranking Method is applied next and finally the clusters are created based on minimum distance between the data point and the centroid. The future scope of this paper is use of Query Redirection can be used to cluster huge amount of data from various databases.

3. METHODOLOGY

In this partition-based robust *K*-means for sequences (*RKMS*) for robust clustering of categorical sequences, the data which is given as input is in character form. There are difficulties in defining an inherently meaningful measure of similarity between sequences. To apply the clustering algorithms and existing indices to sequences, one has to resort to the vectorization method. To convert the sequences in vector form it is encoded into Latin-1, also called ISO-8859-1, is an 8-bit character set endorsed by the International Organization for Standardization (ISO) and represents the alphabets of Western European languages. Once the character file is converted to binary format we are going to use Principal Component Analysis (PCA) algorithm. The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. Using PCA is very important because the dataset which we have it is in very diverse format. Analyzing this type of data becomes very difficult without using PCA. After all this preprocessing we are ready to apply RKMS algorithm on the data.

The number of clusters, *K*, produced by *RKMS* is not necessarily equal to the given number ˆ*K* (we will call ˆ*K* the *expected number* of clusters).The number of clusters is adaptively determined by the elimination of *K*-dependent noise during the *RKMS* clustering. Two key issues for a robust *K*-means type algorithm are: the selection of the initial cluster centroids and the identification of noise clusters. In general, those groups that contain a small number of objects can be viewed as noise. To identify the noise cluster on threshold will be given and if the objects in a cluster are less than threshold then that cluster is considered as noise.

After creating a cluster quality of the cluster is measure by using Cluster Validity Index. For that structural dissimilarity between the cluster object is considered. This new index is linear combination of the within-cluster how objects are scattered and between- cluster how objects are separated and these both measures are computed on the structural dissimilarity of the sequences which is probabilistic approach.

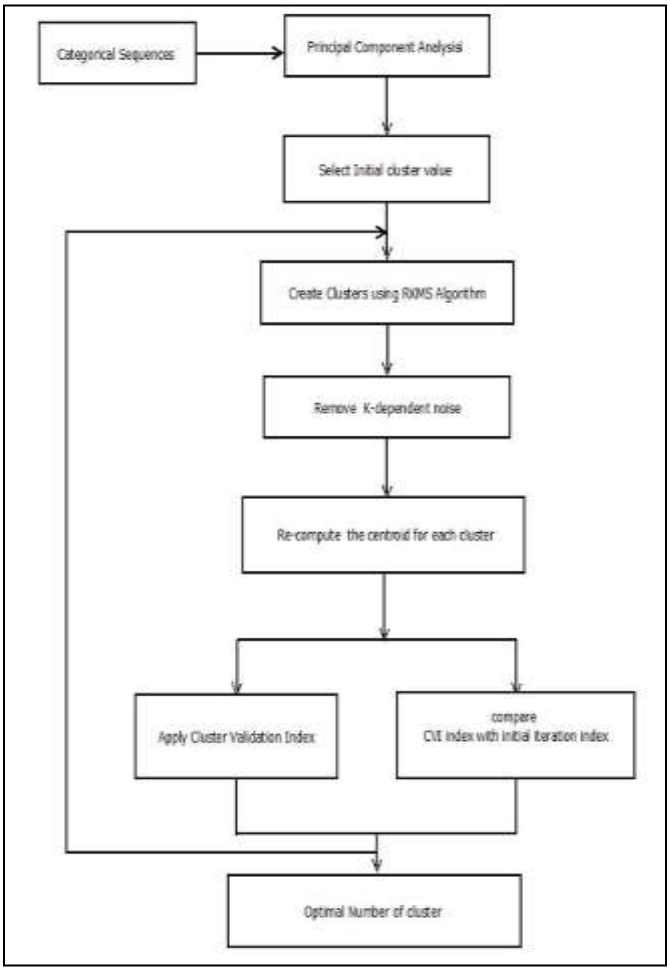
Categorical Sequences
Principal Component Analysis
Select Initial cluster value
Create Clusters using RKMS Algorithm
Remove K-dependent noise
Re-compute the centroid for each cluster
Apply Cluster Validation Index
compare
CVI index with initial iteration index
Optimal Number of cluster

Fig. 1. System Architecture of RKMS method

If we are sign high dimension data then robustness of K-means algorithm may became a problem. There will be two main issues we might face in this algorithm one is selecting the initial centroied value. And another issue will be finding and removing the noise cluster. The first issue raises, because it is known that the performance of the *K*-means algorithm, as an instance of the EM method [13], is sensitive to the initial conditions (here, the initial cluster centroids). Since the number of clusters is fixed at ˆ*K* given by the users, the traditional algorithm likely produces clusters consisting of ˆ*K* -dependent noise, especially when the expected number deviates largely from the true number.

RKMS aims at building a robust condition for the coming iteration, by choosing a set of well-scattered objects (sequences) as the initial cluster centroids. The selection method is based on the greedy approach.

$$J_{\hat{K}}(\mathcal{C}, \mathcal{V}) = \sum_{k=1}^{K} \sum_{s \in C_k} ||\vec{M}(s) - \vec{V}(k)||^2$$

The (k + 1)th initial centroid based on the maximum–minimum principle, given by

$$I_{k+1} = \text{argmax}_{s\in\mathcal{S}\setminus\mathcal{I}_k} \min_{j=1,\ldots,k} ||\vec{M}(s) - \vec{M}(I_j)||^2.$$

In the traditional greedy approach, however, the first centroid is chosen randomly or according to the length of object vector. Since each sequence has been represented in a unit vector as Definition 1 shows,

$$(I_1, I_2) = \text{argmax}_{(s,s')\in\mathcal{S}\times\mathcal{S}} ||\vec{M}(s) - \vec{M}(s')||^2.$$

we choose the first two centroids I1 and I2 by

That is, two sequences having the maximal pairwise distance are chosen as the initial centroids for C1 and C2; the remaining Kˆ − 2 centroids are then determined by repeatedly using (2) with k = 2,..., Kˆ − 1.

RKMS is to delete the possible *K*-dependent noise in each iteration, to further enhance the robustness of the algorithm. The details will be given in Section III-C. Note that the number of resulting clusters produced by *RKMS* could be less than the given *ˆK,* due to the inclusion of this step in the algorithm. This feature is somewhat similar to that of [10] and [12], where the number can be reduced by fading out the redundant clusters; however, *RKMS* targets the identification of *K*-dependent noise.

Given a clustering of the sequence set, *C*, we now aim at establishing a judgement criterion to examine whether each group *Ck* ∈ *C* is a noise cluster or not. In general, those groups that contain a small number of objects can be viewed as noise . According to this view, one can derive a necessary condition for the judgement, i.e., *nk* = |*Ck*| ≤ *τ*, where *nk* is the number of objects in *Ck* and *τ* is a threshold defining the minimal number in a normal cluster. If the condition holds with *τ* = 1, then the cluster is undoubtedly a noise as it consists of one single object. However, it is not the case when 1 < *nk* ≤ *τ* , because the assertion would be concerned with the distribution of the objects in the cluster. To address the problem, we propose a model-selection based method, in view of the observation that if each object in a cluster is indeed *K*-dependent noise, then the preferred clustering model should be the one with that cluster removed. With regard to *Ck* ∈ *C*, we denote the new clustering after the removal of *Ck* by *C_*, created by reassigning each sequence originally belonging to *Ck* to its closest cluster except *Ck*. Thus, the number of clusters satisfies |*C_*| = |*C*| − 1 and, consequently, the conclusion that *Ck* is a noise cluster can be drawn if the model of *C_* is preferred to that of *C*.

Structure-based cluster validity index have two main components first one is structural dissimilarity (SD) of sequences and new cluster validity index(CVI). To measure SD for sequences by Markovian modeling, which is based on the hypothesis that occurrence of each symbol *xl* in the sequence *s* = *x*1 . . . *xl* . . . *xL* is closely related to its preceding subsequence *x*0*x*1 . . . *xl*−1. For a Markov model of order2 □ , the preceding subsequence of *xl* is truncated to *δl* = *xl*−□ . . . *xl*−1, excluding *xl*− □ for *l* < □. By such modeling, the sequential dependence of each symbol *xl* on its preceding subsequence *δl* can be characterized using the conditional probability *ps(xl |δl)* with regard to the sequence *s*.

$$p_s(X) = \sum_{\delta\in\Delta} p_s(\delta) p_s(X|\delta)$$

where *ps(δ)* is the probability of *δ* with regard to *s*.

The SD between two sequences *s* and *s_* can then be numerically computed, by measuring the dissimilarity of their probability distributions *ps(X)* and *ps'(X)*.

$$\text{SD}(s, s') = 1 - \sum_{x\in\mathcal{X}} \sqrt{p_s(x) p_{s'}(x)}.$$

Then next part is CVI *V*SD where it will validate the quality of a series of clustering results, each generated by the clustering algorithm on the same sequence set *S* with various numbers of sequence clusters. To distinguish the results with different cluster numbers, we now use *CK* to denote *C* consisting of *K* sequence clusters, i.e., *CK* = {*C*1, . . . ,*Ck*, . . . ,*CK* }. Denote the minimal and the maximal number of clusters in the series by *K*min and *K*max, respectively. The new index is a linear combination of the within-cluster scatter and between-cluster separation, where both measures are computed on the structural dissimilarity of sequences using SD*(·, ·)* of (5). We define the total within-cluster scatter and between-cluster separation of *CK*, denoted by Scat*(CK)* and Sep*(CK)*, respectively, as

$$\text{Scat}(\mathcal{C}^K) = \sum_{k=1}^{K} \sum_{s\in C_k} \sum_{s'\in C_k} \text{SD}(s, s')$$

$$\text{Sep}(\mathcal{C}^K) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \frac{1}{n_k n_{k'}} \sum_{s\in C_k} \sum_{s'\in C_{k'}} \text{SD}(s, s').$$

Since, in practice, the number of clusters is ranged in [*K*min*, K*max], we define the CVI as a linear combination of the normalized measures, that is

$$V_{\text{SD}}(\mathcal{C}^K) = \frac{\text{Scat}(\mathcal{C}^K)}{\text{maxScat}} + \frac{\text{Sep}(\mathcal{C}^K)}{\text{maxSep}}$$

where maxScat = max*K_*∈[*K*min*,K*max] Scat*(CK_)* and maxSep = max*K_*∈[*K*min*,K*max] Sep*(CK_)*. The index offers a tradeoff between the within-cluster scatter and the between-cluster separation. The minimal *V*SD is considered to be associated with the clustering results produced by the algorithm using the optimal number of clusters.

4. RESULT AND DISSCUSSION

The main purpose is detecting optimal cluster numbers on "USDA Plants Dataset" with using K-means clustering algorithm. Database contains all plants (species and genera) in the database and the states of USA and Canada where they occur.

The dataset used for this research contains 34781 records with 70 features

DataSet Characteristics:	Multivariate
Attribute Characteristics:	Categorical
Associated Tasks:	Clustering
Number of Instances:	22632
Number of Attributes:	70
Missing Values?	Yes

Table -1: Dataset features

To implement this we have Silhouette analysis . It can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

Below we have Silhouette coefficients for the clusters K- means algorithm on 1000 attribute data for the respective values of K.

K- Value Passed	Silhouette coefficients
10	0.161727
20	0.214786
30	0.218644
40	0.255462

Table-2: K values with respective silhouette score

The performance of *RKMS* is evaluated in terms of *clustering accuracy* (*CA*), which is computed as

$$\mathrm{CA} = \frac{1}{N}\sum_{k=1}^{K} a_k$$

where *K* is the true number and *ak* is the number of sequences in the majority group corresponding to *Ck.* To examine the strength of the robust method used to choose the initial cluster centroids in *RKMS*, we also performed clustering using *K*-means with random initialization [1], [4]. The data set was clustered by *K*-means for 20 executions and both the average and best accuracy are reported After applying RKMS algorithm on the same data set the optimal number of cluster we got is as below.

Estimated number of cluster	Silhouette coefficients
23	0.361

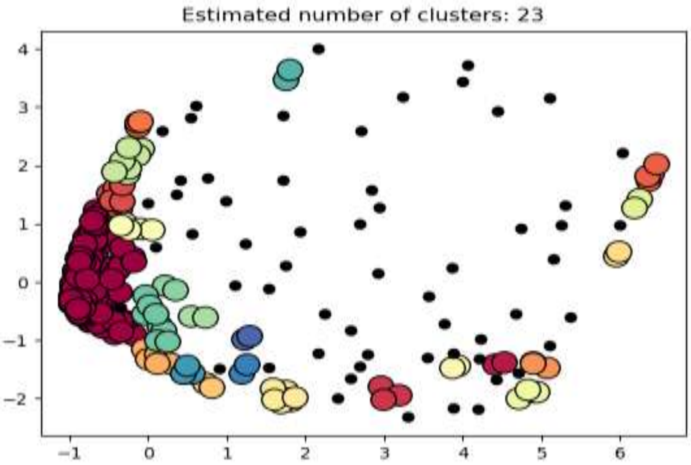

Fig-2. Cluster image with 500 data element

3. CONCLUSION

A robust algorithm named RKMS for partitionbased sequence clustering. In the RKMS algorithm, unlike traditional partitioning methods, the number of clusters (K) can be automatically deduced based on the expected number (ˆK given by the users), by the elimination of cluster-number dependent noise (called K-dependent noise). RKMS and VSD to the common model selection procedure, to estimate the optimal number of clusters in a sequence set. A cluster validation method for categorical sequence clustering is proposed, which is different from the existing methods mainly designed for numeric data. A probabilistic approach to measure the structural dissimilarity of categorical sequences, based on which the within-cluster compactness and between-cluster separation are formulated, in order to evaluate the structural similarity of sequences in each cluster and the structural dissimilarity between sequence clusters. To evaluate the quality of sequence clusters, a new CVI named VSD by linearly combining the within-cluster structural compactness and the between-cluster structural separation.

REFERENCES

1] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering validity checking methods: Part II," *ACM SIGMOD Rec. Arch.*, vol. 31, no. 3, pp. 19–27, 2002.

2] C. C. Aggarwal, *Data Mining: The textbook*. New York, NY, USA: Springer, 2015.

3] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.

4] T. Cali´nski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.

5] H. Sun, S. Wang, and Q. Jiang, "FCM-based model selection algorithms for determining the number of clusters," *Pattern Recognit.*, vol. 37, no. 10, pp. 2027–2037, Oct. 2004.

6] L. Xu, T. W. S. Chow, and E. W. M. Ma, "Topology-based clustering using polar self-organizing map," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 798–808,

7] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proc. 17th ICML*, 2000, pp. 727–734.

8] J. Yang and W. Wang, "CLUSEQ: Efficient and effective sequence clustering," in *Proc. IEEE ICDE*, Mar. 2003, pp. 101–112.

9] M. M.-T. Chiang and B. Mirkin, "Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads," *J. Classif.*, vol. 27, no. 1, pp. 3–40, 2010.

10] G. Dong and J. Pei, "Classification, clustering, features and distances of sequence data," *Seq. Data Min.*, vol. 33, pp. 47–65, 2007.

[11] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

12] K. A. Abdul Nazeer, M. P. Sebastian,.Improving the Accuracy and Efficiency of thek-means Clustering Algorithm.,Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.