A Self Organizing Map Approach for an Enhanced K-means Clustering Algorithm

Dr. M. Sakthi, Associate Professor and Head,

Department of Computer Science, Nallamuthu Gounder Mahalingam College,, Pollachi, TamilNadu

Abstract

The science of analytics has evolved to keep pace with the massive collection of available data. As oceans of data continue to be generated from various sources, analytics plays a very important role to utilize and implement the data. K-means Clustering is an unsupervised data mining technique which makes inferences from datasets using only input vectors without referring to known or labelled outcomes. One of the most important issues in the K-means Clustering Algorithm is the initialization procedure that ultimately determines which part of the solution space will be searched and handling the data with constraints. In order to overcome this narrow span of search with constraints, in this paper, Self Organizing Map (SOM) is used for different initialization procedure with proper training parameters. The proposed algorithm is validated using five different datasets with various performance evaluation metrics. The experimental result shows that the proposed algorithm results in better classification than the standard K-means clustering technique.

Keywords:

K-means Clustering, Self Organizing Map, Data Set, Performance Evaluation Metrics

I. INTRODUCTION

K-means Clustering is an unsupervised clustering algorithm which groups the input data points into multiple classes based on their similarity measures between each other. The grouping is done by minimizing the sum of squared distances between the data points. Two main types of measures used in K-means Clustering are distance measures and similarity measures. The distance measures are used to determine the similarity or dissimilarity of the pair of the objects. Since K-means is the simplest form of clustering, it clusters the data only as the crisp set and it has its own limitations when handling the high dimensional data and the data with constraints. In the real world, with the development of information technology, volumes of data processed by many applications is crossing the peta-scale threshold and so clustering of very large scale data becomes a challenging task nowadays. In this paper, in order to improve the efficiency of K-means Clustering, a Self Organizing Map, which is a form of Artificial Neural Network, is proposed to handle the data with constraints. For effective

validation, the proposed algorithm is applied to five different datasets namely Iris, Wine, Lung cancer, Yeast and Glass. The performance metrics used for validating the proposed algorithm are Precision and Recall, Error Rate and Accuracy and Execution Time. The experimental result shows that the proposed algorithm offers better opportunity for an early exploration of the search space, and as the process continues it gradually narrows the search for constrained data.

II. RELATED WORKS

Trujillo et al., [5] proposed a combining K-means based grid clustering approach. Clustering is widely used in various applications which include data mining, information retrieval, image segmentation, and data classification. A clustering technique for grouping data sets that are indexed in the space is proposed in this paper. This approach mainly depends on the k-means clustering technique and grid clustering. K-means clustering is the simplest and most widely used approach. The main disadvantage of this approach is that it is sensitive to the selection of the initial partition. Grid clustering is extensively used for grouping data that are indexed in the space. The main aim of the proposed clustering approach is to eliminate the high sensitivity of the k-means clustering approach to the starting conditions by using the available spatial information. A semivariogram based grid clustering technique is used in this approach. It utilizes the spatial correlation for obtaining the bin size. The author combines this approach with a conventional k-means clustering technique as the bins are constrained to regular blocks while the spatial distribution of objects is irregular. An effective initialization of the k-means is provided by semivariogram. From the experimental results, it is clearly observed that the final partition protects the spatial distribution of the objects.

Yanfeng Zhang et al., [3] proposed an Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number (NSS-AKmeans) approach for learning optimal number of clusters and for providing significant clustering results. High density areas can be detected by the NSS-AKmeans and from these centers the initial cluster centers with a neighbour sharing selection approach can also be determined. Agglomeration Energy (AE) factor is proposed in order to choose a initial cluster for representing global density relationship of objects. Moreover, in order to calculate local neighbour sharing relationship of objects, Neighbors Sharing Factor (NSF) is used. Agglomerative Fuzzy K-means clustering algorithm is then utilized to further merge these initial centers to get the preferred number of clusters and create better clustering results. Experimental observations on several data sets have proved that the proposed clustering approach was very significant in automatically identifying the true cluster number and also providing correct clustering results

Zhang Zhe et al., [1] proposed an improved K-means clustering algorithm. K-means algorithm [8] is extensively utilized in spatial clustering. The mean value of each cluster centroid in this approach is taken as the Heuristic information, so it has some limitations such as sensitive to the initial centroid and instability. The enhanced clustering algorithm referred to the best clustering centroid which is searched during the optimization of clustering centroid. This increases the searching probability around the best centroid and enhanced the strength of the approach. The experiment is performed on two groups of representative dataset and from the experimental observation, it is clearly noted that the improved K-means algorithm performs better in global searching and is less sensitive to the initial centroid

Huang et al., [6] put forth the automated variable weighting in k-means type clustering that can automatically estimate variable weights. A novel approach is introduced to the K-means algorithm to iteratively update variable weights depending on the present partition of data and a formula for weight calculation is also proposed in this paper. The convergency theorem of the new clustering algorithm is given in this paper. The variable weights created by the approach estimates the significance of variables in clustering and can be deployed in variable selection in various data mining applications where large and complex real data are often used. Experiments are conducted on both synthetic and real data and it is found from the experimental observation that the proposed approach provides higher performance when compared the traditional k-means type algorithms in recovering clusters in data.

In order to support the visual analysis of spatiotemporal data, Andrienko*et al.,* (2010) suggest a framework based on the "Self-Organizing Map" method combined with a set of interactive visual tools supporting both analytic perspectives. SOM can be considered as a combination of clustering and dimensionality reduction. In the first perspective, SOM is applied to the spatial situations at different time moments or intervals. In the other perspective, SOM is applied to the local temporal evolution profiles. The integrated visual analytics environment includes interactive coordinated displays enabling various transformations of spatiotemporal data and post-processing of SOM results. The SOM matrix display offers an overview of the groupings of data objects and their two-dimensional arrangement by similarity. This view is linked to a cartoFigureic map display, a time series Figure, and a periodic pattern view. The linkage of these views supports the analysis of SOM results in both the spatial and temporal contexts.

Chattopadhyay*et al.,* (2012) proposed a visual clustering approach for machine part cell formation using self organizing map algorithm, an unsupervised neural network to achieve

better group technology efficiency measure of cell formation as well as measure of SOM quality. The work also has established the criteria of choosing an optimum SOM size based on results of quantization error, and average distortion measure during SOM training which have generated the best clustering and preservation of topology. To evaluate the performance of the proposed algorithm, this work has tested the several benchmark problems available in the literature.

Ecological data are considered to be difficult to analyse because numerous biological and environmental factors are involved in a complex manner in environment organism relationships. The Self-Organizing Map has advantages for information extraction (i.e., without prior knowledge) and the efficiency of presentation (i.e., visualization). It has been implemented broadly in ecological sciences across different hierarchical levels of life. Recent applications of the SOM, which are reviewed here, include the molecular, organism, population, community, and ecosystem scales. Further development of the SOM is discussed regarding network architecture, spatio-temporal patterning, and the presentation of model results in ecological sciences are given in Chon and Tae-Soo (2011).

III. METHODOLOGY

In this paper, in order to improve the efficiency of K-means Clustering, a Self Organizing Map, which is a form of Artificial Neural Network, is proposed to handle the data with constraints

K-means Clustering using Self Organizing Map

The Self-Organizing Map (SOM) is a clustering and data visualization technique based on neural networks. SOM can be viewed as a vector quantization technique where the reference vectors are learned by training a neural network. Thus, like vector quantization, the goal of SOM is to find a set of reference vectors and to assign each point in the data set to the “closest” reference vector. With the SOM approach, each reference vector is associated with a particular neuron, and the components of that reference vector become the “weights” of that neuron. As with other neural network systems, a SOM neural net is trained by considering the data points one at a time and adjusting the weights (reference vectors) so as to best fit the data.

The final output of the SOM technique is a set of reference vectors which implicitly defines clusters. (Each cluster consists of the points closest to a particular reference vector.) However, while SOM might seem very similar to K-means or other vector quantization

approaches, there is a fundamental conceptual difference. During the training process, SOM uses each data point to update the closest reference vector and the reference vectors of nearby neurons. In this way, SOM produces an "ordered" set of reference vectors. In other words, the reference vectors of neurons which are close to each other in the SOM neural net will be more closely related to each other than to the reference vectors of neurons that are farther away in the neural net (Ismail et al., 2011). Because of this constraint, the reference vectors in SOM can be viewed as lying on a smooth, elastic two-dimensional surface in *m* dimensional space, a surface which tries to fit the *m* dimensional data as well as possible.

The basic idea of a SOM is to map the data patterns onto a*n*-dimensional grid of neurons or units. That grid forms what is known as the output space, as opposed to the input space where the data patterns are. This mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space, and vice-versa. So as to allow an easy visualization, the output space is usually 1 or 2 dimensional.

The proposed SOM training algorithm can be described as follows:

Let X be the set of n training patterns $\mathbf{x_1, x_2,..,x_n}$

W be a p×q grid of units $\mathbf{w_{ij}}$where i and j are their coordinates on that grid

⟨ be the learning rate, assuming values in [0,1], initialized to a given initial learning rate

r be the radius of the neighbourhood function $\mathbf{h(w_{ij},w_{mn},r)}$,initialized to a given initial radius

1 Repeat

2 For k=1 to n

3 For all $\mathbf{w_{ij}} \in W$, calculate $d_{ij} = \| \mathbf{x_k - w_{ij}} \|$

4 Select the unit that minimizes d_{ij} as the winner $\mathbf{w_{winner}}$

5 Update each unit $\mathbf{w_{ij}} \in W$: $w_{ij} = w_{ij} + \langle h(w_{winner},w_{ij},r) \| x_k - w_{ij} \|$

6 Decrease the value of ⟨ and r

7 Until ⟨ reaches 0

The neighbourhood function *h* is usually a function that decreases with the distance (in the output space) to the winning unit, and is responsible for the interactions between different units. During training, the radius of this function will usually decrease, so that each unit will become more isolated from the effects of its neighbours. It is important to note that many implementations of SOM decrease this radius to 1, meaning that even in the final stages of

training; each unit will have an effect on its nearest neighbours, while other implementations allow this parameter to decrease to zero.

SOMs can be used in many different ways, even within clustering tasks. In this work it is assumed that each SOM unit is a cluster centre, and thus a k-unit SOM will perform a task similar to K-means. It must be noted that SOM and K-means algorithms are rigorously identical when the radius of the neighbourhood function in the SOM equals zero. In this case the update occurs only in the winning unit just as happens in K-means (step 4).

Self Organizing Maps are iterative algorithms that discover a set of representatives. The SOM algorithm for K-means clustering is a special case of a SOM; in this case, the representatives tend toward the means of clusters defined by a popular squared error criterion. Interpretation is complicated by the difficulty of identifying a specified SOM with an error criterion that it might be considered to minimize. The derivation of the SOM algorithm for K-means clustering appears to be a special case analysis of SOMs.

In general, however, SOMs cannot be interpreted as algorithms that minimize a fixed error criterion to the process of discretizingR^p. Here each bin is represented by averaging the xi that lie within it. SOMs offer a compromise between discovering structure in the data (clustering the x_i) and imposing structure upon the data (discretizingR^p).

The signature feature of SOMs is the imposition of an external topological ordering on the set of representatives. This is usually accomplished by identifying the representatives with a regular grid in one or two dimensions. The topology of the grid is then used to define neighbourhoods of representatives. As the idea is that the representatives (called `weights') are spatially correlated, so that representatives at nearby points on the grid are more similar than those which are widely separated. Generalizing the SOM algorithm for K-means clustering, "on-line" SOMs update all of the representatives in a neighbourhood of $m_*(x)$. This work illustrates this kind of methodology, and explores some of its implications, with a trivial example.

IV. EXPERIMENTAL RESULTS

The datasets used in this work are taken from the UCI Machine Learning Repository which is a collection of databases, domain theories, and data generators used by the machine learning community for the empirical analysis of machine learning algorithms. In this work,

iris dataset, wine dataset, lung cancer dataset, glass dataset and yeast dataset are used for the experiment and the performance is evaluated based on accuracy and execution of the algorithm, precision and recall measures and the error rate metrics. These three metrics are very important to evaluate the performance and efficiency of any algorithms. The performance of the K-means is based on the clustering with Self Organizing Maps on these datasets. Vectors, dimensions and clusters for all the datasets are shown in Table 4.1.

Table 4.1 Vectors, Dimensions and clustering of the datasets

Datasets	Vectors	Dimensions	Clusters
Iris	150	4	3
Wine	178	13	3
Lung cancer	32	56	5
Glass	214	9	7
Yeast	1484	8	10

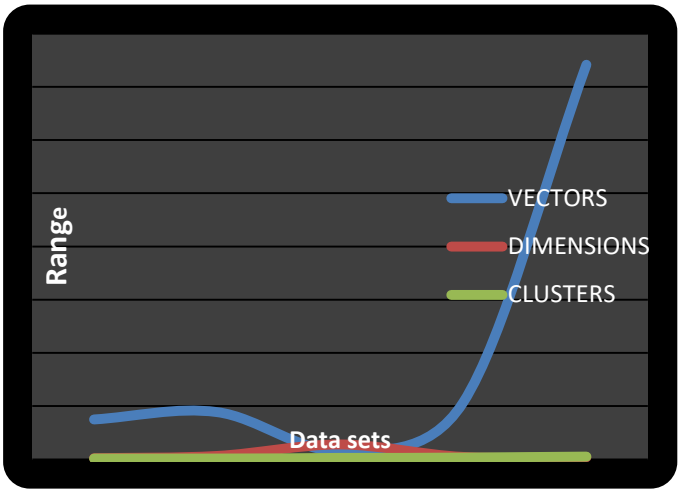

Figure 4.1 Training Datasets with Vectors, Dimensions and Clusters

4.1 Performance Evaluation Measure

4.1.1 Accuracy and Execution Time

Table 4.2 shows the accuracy and execution time of K-means clustering and

K-means with SOM with five datasets.

Table 4.2 Accuracy and Execution Time for SOM

Datasets	K-means		K-means with SOM	
	Accuracy (%)	Execution Time (Seconds)	Accuracy (%)	Execution Time (Seconds)
Iris	56	38	78	32
Wine	62	34	81	29
Lung cancer	77	30	85	26
Glass	79	28	89	23
Yeast	83	23	93	16

Figure 4.2 shows the accuracy of K-means clustering with SOM. The comparison of K-means clustering and K-means with SOM proves that the values 78, 81, 85, 89 and 93 of SOM has high accuracy in proposed method when compared with the K-means clustering algorithm

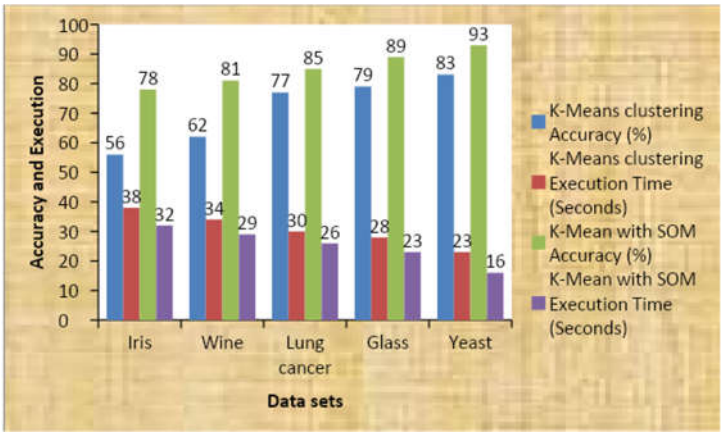

Figure 4.2 Accuracy and Execution time of K- Means and SOM

4.1.2 Precision and recall

The precision percentage and Recall percentage of Self Organizing Map is tabulated in Table 4.3

Table 4.3 Precision and Recall Percentage of SOM

Datasets	K-means Clustering		K-Mean with SOM	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Iris	63	73	76	70
Wine	68	71	79	69
Lung cancer	74	68	76	65
Glass	79	63	81	58
Yeast	83	54	88	43

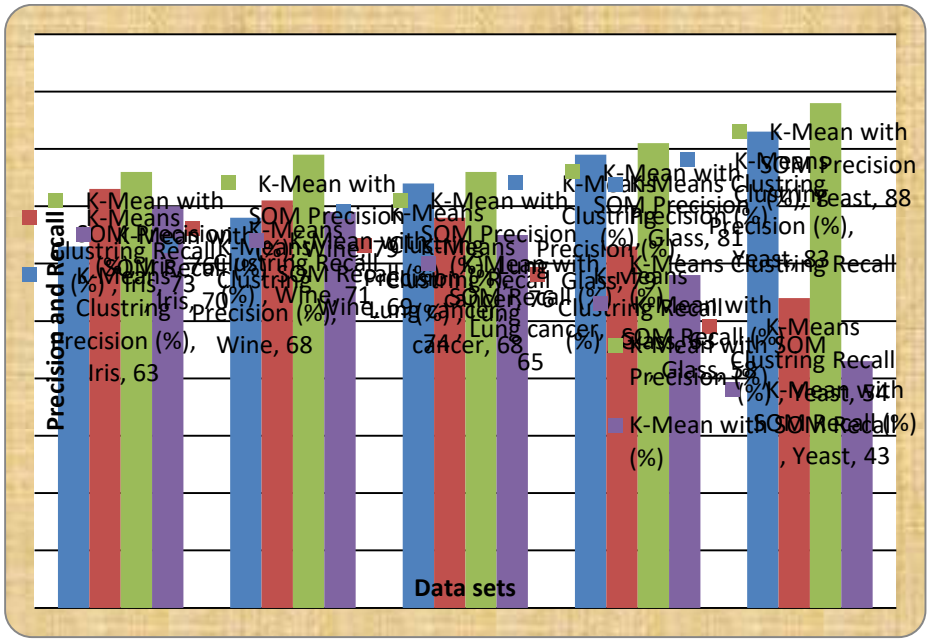

Figure 4.3 Precision and Recall Percentage of K-means and K-means with SOM

4.1.3 Error Rate

In Table 4.4 the error rate in percentage is described for calculating the K-Mean clusters using various datasets are Iris, Wine, Lung Cancer, Glass and Yeast.

Table 4.4 Error Rate of SOM

Datasets	Error Rate (%)	
	K-means clustering	K-Mean with SOM
Iris	36	24
Wine	27	21
Lung cancer	21	19
Glass	19	12
Yeast	11	8

The error rate for clustering techniques is shown in Figure 4.4. Error rate is very less in proposed SOM technique than K-means approach and thus it is proved that the K-means with SOM has better performance than the K-means Clustering Algorithm.

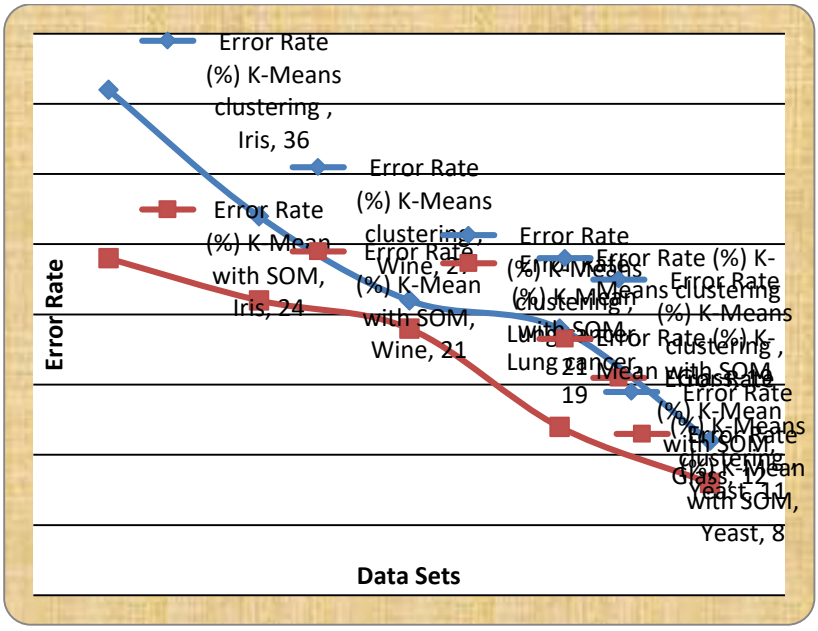

Figure 4.4 Error Rate of K-means and K-means with SOM

V. CONCLUSION

The need of analyzing and grouping of data is required for better understanding and examination of data. This can be solved by using the clustering technique which groups the similar kind of data into a particular cluster. In K-means Clustering, the initial centroid is

generated randomly before clustering. If the dataset used is large, then the performance of K-means will be reduced and also the time complexity is increased. The major limitation of K-Means Clustering Algorithm is the initialization procedure that ultimately determines which part of the solution space will be searched and handling the data with constraints. In order to overcome this, the Self Organizing Map(SOM) is used for different initialization procedures with proper training parameters. The proposed algorithm is validated using five different datasets with various performance evaluation metrics. The experimental result shows that the proposed algorithm results in better classification than the standard K-means clustering algorithm.

VI. REFERENCES

[1] Zhang Zhe, Zhang Junxi and Xue Huifeng, "Improved K-means Clustering Algorithm", Congress on Image and Signal Processing, Vol. 5, Pp. 169-172, 2008.

[2] Hai-xiang Guo, Ke-jun Zhu, Si-wei Gao and Ting Liu, "An Improved Genetic k-means Algorithm for Optimal Clustering", Sixth IEEE International Conference on Data Mining Workshops, Pp. 793-797, 2006.

[3] Yanfeng Zhang, Xiaofei Xu and Yunming Ye, "NSS-AKmeans: An Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number", 2nd International Conference on Advanced Computer Control, Vol. 2, Pp. 32-38, 2010.

[4] Xiaoyun Chen, Youli Su, Yi Chen and Guohua Liu, "GK-means: an Efficient K-means Clustering Algorithm Based on Grid", International Symposium on Computer Network and Multimedia Technology, Pp. 1- 4, 2009.

[5] Trujillo, M., Izquierdo, E., "Combining K-means and semivariogrambased grid clustering", 47th International Symposium, Pp. 9-12, 2005.

[6] Huang, J.Z., Ng, M.K., Hongqiang Rong and Zichen Li, "Automated variable weighting in k-means type clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 5, Pp. 657-668, 2005.

[7] Yi Hong and Sam Kwong “Learning Assignment Order of Instances for the constrained k-means clustering algorithm” IEEE Transactions on Systems, Man, and Cybernetics, Vol 39, No 2. April, 2009.

[8] I. Davidson,M. Ester and S.S. Ravi, “Agglomerative hierarchical clustering with constraints: Theoretical and empirical results”, in Proc. of Principles of Knowledge Discovery from Databases, PKDD 2005.

[9] Wagstaff, Kiri L., Basu, Sugato, Davidson, Ian “When is constrained clustering beneficial, and why?” National Conference on Artificial Intelligence, Boston, Massachusetts 2006.

[10] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrodl “Constrained K-means Clustering with Background Knowledge” ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001.

[11] I. Davidson, M. Ester and S.S. Ravi, “Efficient incremental constrained clustering”. In Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, August 12-15, San Jose, California, USA.

[12] I. Davidson, M. Ester and S.S. Ravi, “Clustering with constraints: Feasibility issues and the K-means algorithm”, in proc. SIAM SDM 2005, Newport Beach, USA.

[13] D. Klein, S.D. Kamvar and C.D. Manning, “From Instance-Level constraints to space-level constraints: Making the most of Prior Knowledge in Data Clustering”, in proc. 19th Intl. on Machine Learning (ICML 2002), Sydney, Australia, July 2002, p. 307-314.

[14] N. Nguyen and R. Caruana, “Improving classification with pairwise constraints: A margin-based approach”, in proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD’08).

[15] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, “Constrained Kmeans clustering with background knowledge”, in: Proc. Of 18th Int. Conf. on Machine Learning ICML’01, p. 577 - 584.

[16] Y. Hu, J. Wang, N. Yu and X.-S. Hua, “Maximum Margin Clustering with Pairwise Constraints”, in proc. of the Eighth IEEE International Conference on Data Mining (ICDM) , 253-262, 2008.