ISSN 22773061

Comparative Analysis of Kohonen-SOM and K-Means data mining algorithms based on Academic Activities

Shaina Dhingra
Lovely Professional University
Jalandhar, India

Rimple Gilhotra
Lovely Professional University
Jalandhar, India

Ravishanker
Lovely professional University
Jalandhar, India

shaina.dhingra@gmail.com

ABSTRACT

With the increasing demand of IT and subsequent growth in this sector, the high- dimensional data came into existence. Data Mining plays an important role in analyzing and extracting the useful information. The key information which is extracted from a huge pool of data is useful for decision makers. Clustering, one of the techniques of data mining is the mostly used methods of analyzing the data. In this paper, the approach of Kohonen SOM and K-Means and HAC are discussed. After that these three methods are used for analyzing the academic data set of the faculty members of particular university. Finally a comparative analysis of these algorithms are done against some parameters like number of clusters, error rate and accessing rate, etc. This work will present new and improved results from large-scale datasets.

Keywords

Kohonen- SOM, K-means, HAC, PCA

Council for Innovative Research
Peer Review Research Publishing System
Journal: ***INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY***
Vol 6, No 1

editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

INTRODUCTION

A self-organizing map (SOM) or self-organizing feature map (SOFM) is a kind of artificial neural network uses unsupervised learning to produce a low-dimensional (typically two dimensional), discretized representation of the input space of the training samples, called a map. SOM differs from other artificial neural networks because it uses a neighborhood function to preserve the topological properties of the input space.

SOM is one of the clustering methods. SOM organizes the data in clusters (cells of map) such as the values in the same cell are similar, and the values in different cells are different. According to this, SOM gives comparable results to state-of-the-art clustering algorithm such as K-Means.

SOM is also used as a visualization technique. It allows us to visualize in a low dimensional [2] representation space (2D) the original dataset. Indeed, the individuals located in adjacent cells are more similar than individuals located in distant cells. In this point of view, it is comparable to visualization techniques such as Multidimensional scaling or PCA (Principal Component Analysis).

Through this, it can be showed how to implement the Kohonen's SOM algorithm with a particular tool. After implementation it has been tried to assess the properties of this approach by comparing the results with those of the PCA algorithm. Then, compare the results to those of K-Means. Finally [3] the Two-step Clustering process is implemented by combining the SOM algorithm with the HAC process (Hierarchical Agglomerative Clustering). It is a variant of the Two-Step clustering where combine K-Means and HAC.

The Kohonen algorithm is a very powerful tool for data analysis. It was originally designed to model organized connections between some biological neural networks. It was also immediately considered as a very good algorithm to realize vectorial quantization, and at the same time pertinent classification, with nice properties for visualization. If the individuals are described by quantitative variables (ratios, frequencies, measurements, amounts, etc.), the straightforward application of the original algorithm leads to build code vectors and to associate to each of them the class of all the individuals which are more similar to this code-vector than to the others. But, in case of individuals described by categorical (qualitative) variables having a

finite number of modalities (like in a survey), it is necessary to define a specific algorithm. In this paper, we present a new algorithm inspired by the SOM algorithm, which provides a simultaneous classification of the individuals and of their modalities.

RELATED WORK

Ji Dan, Qiu Jianlin in [14]-In this paper the author has discussed the two techniques of Data Mining, namely, clustering and decision trees. The main focus of this paper is to present a new Data Mining algorithm called CA which improves the actual methods of CURE and c4.5 algorithms. The new algorithm introduces the PCA, grid partition and parallel processing which are used for the feature reduction and scale reduction for high dimensional data set. The author has also discussed the different types of clustering and the algorithms those come under different categories of clustering. In addition to this, the Classification technique, namely decision Tree, is also included in this paper. The CA algorithm is presented as a combination of PCA, CURE and C4.5 algorithms. The main procedure of CA algorithms includes: Feature reduction, Scale reduction and Decision Tree Classification. Finally, the CA algorithm is applied in maize seed breeding to find some useful information for seed breeding. According to the author the efficiency of CA is higher not only in clustering but also in decision tree.

Timothy C. Havens et James C. Bezdek in [15]-The main concern of this paper is to use the clustering for Very Large data sets. The author compares the methods based on sampling, incremental techniques and kernalized version of FCM. Numerical experiments are performed on both the laudable and Very large data sets. These experiments enables the comparisons based on time and space complexity, speed, etc. The main focus of this paper is to compare four algorithms of fuzzy partitions of very large data sets, namely, rseFCM,spFCM, okFCM, brFCM and to discuss the FCM algorithms for Very large data set. Additionally, a new extension of some of these algorithms is purposed. Finally, two experiments are performed. One experiment is done to compare the performance of Very large FCM algorithms on data and in the other set of experiments, the purposed algorithms are applied to data sets.

METHODOLOGY

Kohonen-SOM's approach

Kohonen's SOM is called a topology-preserving map because there is a topological structure imposed on the nodes in the network. A topological map is simply a mapping that preserves neighborhood relations.

Algorithm for Kohonen's Self Organizing Map

- Assume output nodes are connected in an array (usually 1 or 2 dimensional)
- Assume that the network is fully connected - all nodes in input layer are connected to all nodes in output layer.
- Use the competitive learning algorithm as follows:

1. Randomly choose an input vector x

2. Determine the "winning" output node i, where wi is the weight vector connecting the inputs to output node i. The above equation is equivalent to

w_i x >= w_k x only if the weights are normalized.

$$|\omega_i\text{-}X| \leq |\omega_K\text{-}X| \quad \forall k$$

Given the winning node i, the weight update is

ISSN 22773061

$\omega_k(new) = \omega_k(old) + \Delta\omega_k(n)$

where $\Delta\omega_k(n)$ represesnts the change in weight**.**

K-Means's approach

The **K-Means** algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the 'inertia' of the groups. This algorithm requires the number of cluster to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields. It is also equivalent to the expectation-maximization algorithm when setting the covariance matrix to be diagonal, equal and small. The K-means algorithm aims to choose centroids C that minimize the within cluster sum of squares objective function with a dataset X with n samples:

$J(X,C) = \sum_{i=0}^{n} \min \left(||X_j-\mu_i||^2 \right)$ where, $\mu_i \varepsilon C$

K-means is often referred to as Lloyd's algorithm. In basic terms, the algorithm has three steps. The first step chooses the initial centroids, with the most basic method being to choose k samples from the dataset X. After initialization, k-means consists of looping between the other two major steps. The first step assigns each sample to its nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids is the inertia and the algorithm repeats these last two steps until this value is less than a threshold.

RESULTS & DISCUSSION

Data Set

I am using "faculty activities" dataset it has real values there are approximately 20 descriptors and 3303 instances. Some of them will be used as an input attributes and other are used as an output attributes.

Table 1: Attributes of Data Set

QOS	Qualilty of Syllabus
QOIP	Quality of Instruction Plan
HOCC	Handling of committees clubs
QOIP	Qualilty of Instruction Plans
RM	Record Maintenance
EVA	Evaluation
QP	Question Ppaer
CACP	CA of capstone project
RPP	Research Paper published
CA	Confrence atttended paper published
WA	Workshop attended
BP	Book Cumputers published
DM	Discipline Maintenance
IWS	Interactivity with students
EMTE	Evaluation of MTE and ETE
AWA	Award
LA	Leadership ability
FDP	Faculty Development program
QUAL	Qualification
Grading	Grading

The easiest way to import a XLS file in Tanagra is to create a new file in Tanagra and import the XLS file.

In the figure 1, the grid with 2 rows and 3 columns i.e. 2*3=6 clusters has been created in Tanagra. After setting the parameters for Kohonen SOM, the above shown output appears on screen that is in the form of MAP Topology and MAP Quality. To improve the visualization the PCA is implemented on Kohonen SOM. The figure 3 depicts the different clusters which are created by applying the K-Means algorithm on the academic data set. The figure 4 depicts the different clusters which are created by applying the K-Means algorithm and PCA on the academic data set.

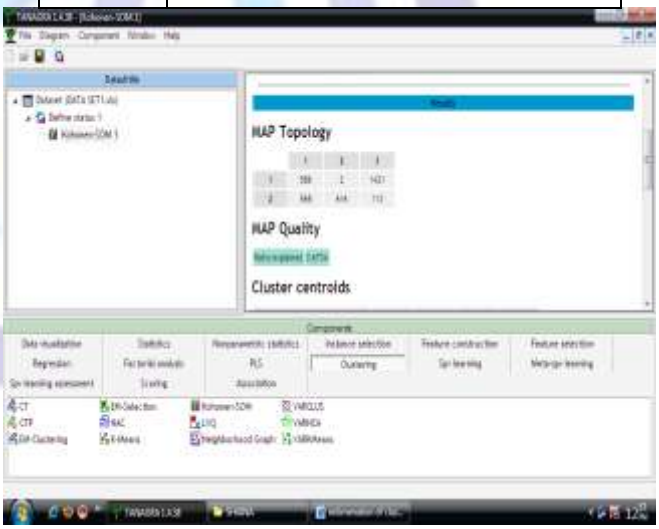

Figure 1. Implemented Kohonen SOM Algorithm and generate MAP Topology with 6 clusters, MAP Quality 0.8189

ISSN 22773061

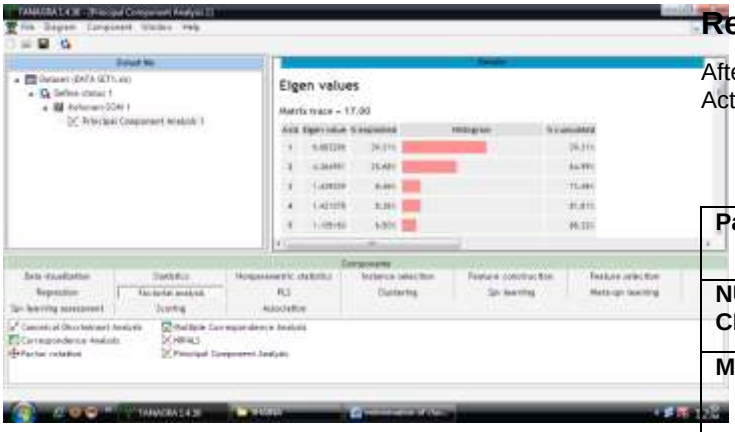

Figure 2. Implemented PCA on Kohonen SOM and evaluate Eigen Values

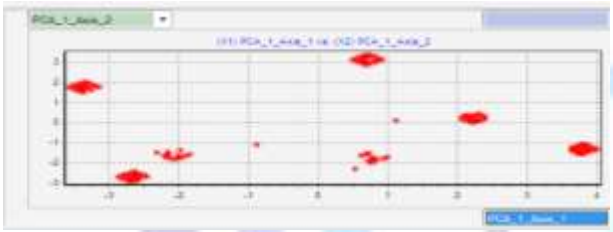

Figure 3. Now K-Means Showing 6 different clusters with same scale.

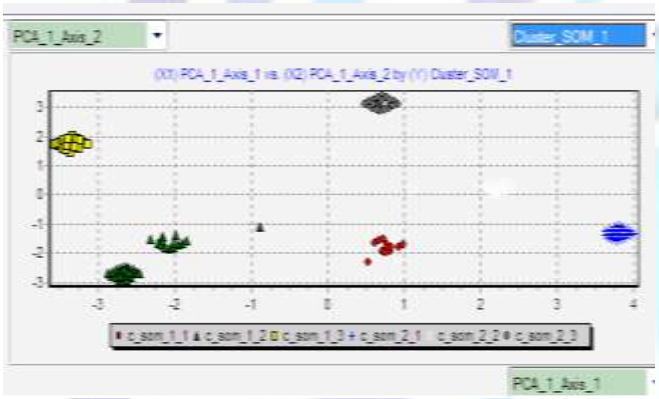

Figure 4. The KOHONEN-SOM component shows 6 different clusters with different scale

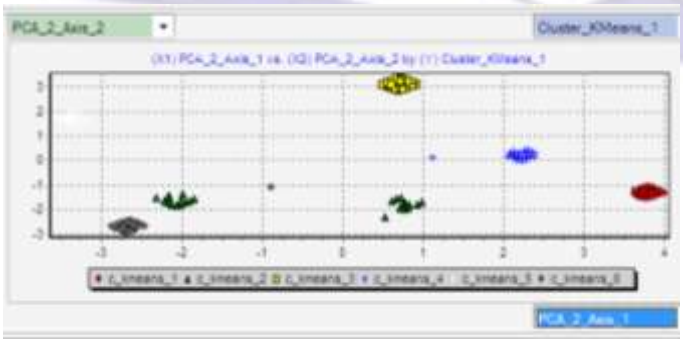

Figure 5. The K-Means algorithm shows 6 different clusters with different Conclusion

The above figure represents the improved visualization after applying the PCA on K-means.

Result

After implementation of these algorithms on Academic Activities data set, the following results obtained:

Table 2: Comparative results of both algorithms

Parameters	Kohonen SOM	K-MEANS
NUMBER OF CLUSTERS	6	6
MAP TOLOPLOGY	6	6
ERROR RATE	0.8189	0.8456
COMPUTAION TIME	297 MS	1281 MS
ACEESING TIME	FAST	SLOW

According to the figures, Kohonen Som and K-Means have a same number of clusters and Map Topology. After implementation, it is analyzed that the Kohonen Som produced better results in terms of error rate, computation time and access time. In the next step I add the PRINCIPAL COMPONENT ANALYSIS (PCA) under view dataset 1 component. This component is added to visualize the dataset in lower dimensions. It can be seen as dimensionality reduction technique. Without PCA we cannot clearly visualize the data, clusters are overridden on each other. PCA is a way of identifying patterns in data and expressing data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in the data of high dimension, where luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The main advantage of PCA is that once you have found these patterns in the data and you compress the data i.e. by reducing number of dimensions, without much loss of information. The next screenshot is displaying the output while after adding PCA component.

CONCLUSION

This comparative study of two algorithms for the clustering, namely, Kohonen SOM and K-Means suggest that Kohonen SOM gives better performance as compare to K-means.the performance of these two algorithms is measured on the basis of few parameters like number of clusters, map topology, error rate, computation time and accessing time. From this comparative study, it is concluded that (i) Kohonen-SOM comparatively gives less error rate or high accuracy, (ii) Computation time taken by the Kohonen SOM is very less as compare to the time taken by K-Means on the same data set. Finally, it can be stated, when tested in a completely equal working conditions, Kohonen SOM can be considered as an appropriate clustering algorithm for high dimensional data set.

REFERENCES

[1] Rakesh Agrawal , Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules in Large

Databases, Proceedings of the 20th International Conference on Very Large Data Bases, p.487-499, September 12-15, 1994

[2] A. Baskurt, F. Blum, M. Daoudi, J.L. Dugelay, F. Dupont, A. Dutartre, T. Filali Ansary, F. Fratani, E. Garcia, G. Lavoué, D. Lichau, F. Preteux, J. Ricard, B. Savage, J.P. Vandeborre, T. Zaharia. SEMANTIC-3D : COMPRESSION, INDEXATION ET TATOUAGE DE DONNÉES 3D Réseau National de Recherche en Télécommunications (RNRT) (2002)

[3] T.Zaharia F.Prêteux, Descripteurs de forme : Etude comparée des approches 3D et 2D/3D 3D versus 2D/3D Shape Descriptors: A Comparative study

[4] T.F.Ansary J.P.Vandeborre M.Daoudi, Recherche de modèles 3D de pièces mécaniques basée sur les moments de Zernike

[5] A. Khothanzad, Y. H. Hong, Invariant image recognition by Zernike moments, IEEE Trans. Pattern Anal. Match. Intell.,12 (5), 489-497, 1990.

[6] Agrawal R., Imielinski T., Swani A. (1993) Mining Association rules between sets of items in large databases. In : Proceedings of the ACMSIGMOD Conference on Management of Data, Washington DC, USA.

[7] Agrawal R., Srikant R., Fast algorithms for mining association rules in larges databases. In Proceeding of the 20th international conference on Very Large Dada Bases (VLDB'94),pages 478-499. Morgan Kaufmann, September 1994.

[8] U. Fayyad, G.Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining toKnowledge Discovery in Databases, American Association for Artificial Intelligence. All rights reserved. 0738-4602-1996

[9] S.Lallich, O.Teytaud, Évaluation et validation de l'intérêt des règles d'association

[10] Osada, R., Funkhouser, T., Chazelle, B. et Dobkin, D. ((Matching 3D Models with Shape Distributions)). Dans Proceedings of the International Conference on Shape Modeling & Applications (SMI '01), pages 154–168. IEEE Computer Society,Washington, DC, Etat-Unis. 2001.

[11] W.Y. Kim et Y.S. Kim. A region-based shape descriptor using Zernike moments. Signal Processing : Image Communication, 16 :95–100, 2000.

[12] A. Khotanzad et Y.H. Hong. Invariant image recognition by Zernike moments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(5), May 90.

[13] N Pasquier, Y Bastide, R Taouil, L Lakhal - Database Theory ICDT'99, 1999 – Springer

[14] Ji Dan, Qiu Jianlin et Gu Xiang, Chen Li, He Peng. (2010) A Synthesized Data Mining Algorithm based on Clustering and Decision tree. 10th IEEE International Conference on Computer and Information Technology (CIT 2010)

[15] Timothy C. Havens et James C. Bezdek. Fuzzy c-Means Algorithms for Very Large Data