Volume 4, Issue 10, October 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Data Mining Clustering Approach for Traffic Accident Analysis of National Highway-1

Manisha Birdi[1], Prof. Dr. R C Gangwar[2], Prof. Gurpreet Singh[3]
[1]M.Tech*(CSE), Beant College of Engineering And Technology, Gurdaspur, India
[2]associate Professor/Head CSE, Beant College of Engineering And Technology, Gurdaspur, India
[3]Research Scholar (Ph.D Computer Sc.), Pacific Academy of Higher Education & Research University, Udaipur, India

Abstract*- Classifying the main causative aspects to traffic accidents and their harshness will assist highway safety development ingenuities by better capability design and learning program to address the needs due to the changes in demographics. The traffic accidents data used in this study has been together over the last 6 months on the rural highways and urban streets from NH-1, Delhi-Amritsar (India). In order to determine the major factors contributing to traffic accidents and their severity, we present a data mining Algorithms using K-Means and SOM clustering algorithms to analyze the traffic accidents data. The experiment results from this study will show that the established data mining model using clustering can successfully classify the major contributing factors to traffic accidents and their collision severity for different groups of people with good accuracy. The data mining model is evaluated and compared with a commercial software package Tanagra. Recommendations drawn from the study results for traffic safety improvements are presented.

***Index Terms*— Clustering , data mining, traffic collision, K-Mean and SOM algorithms, PCA.**

I. INTRODUCTION

A self-organizing map (SOM) or self-organizing feature map (SOFM) is a kind of artificial neural network that [1] is trained using unsupervised learning to produce a low-dimensional (typically two dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different than other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space.

SOM is a clustering method. Indeed, it organizes the data in clusters (cells of map) such as the instances in the same cell are similar, and the instances in different cells are different. In this point of view, 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, which is a clustering algorithm. Finally implement [3] the Two-step Clustering process 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.

II. RELATED WORK

According to Ji Dan, Qiu Jianlin [14], with the development of information technology and computer science, high-capacity data appear in our lives . In order to help people analyzing and digging out useful information, the generation and application of data mining technology seem so significance. Clustering and decision tree are the mostly used methods of data mining. Clustering can be used for describing and decision tree can be applied to analyzing. After combining these two methods effectively, it can reflect data characters and potential rules syllabify. This paper presents a new synthesized data mining algorithm named CA which improves the original methods of CURE and C4.5. CA

introduces principle component analysis (PCA), grid partition and parallel processing which can achieve feature reduction and scale reduction for large-scale datasets. This paper applies CA algorithm to maize seed breeding and the results of experiments show that our approach is better than original methods.
According to Timothy C. Havens et James C. Bezdek [15], very large (VL) data or "Big Data" are any data that you cannot load into your computer's working memory. This is not an objective definition, but a definition that is easy to understand and one that is practical, because there is a data set too big for any computer you might use; hence, this is VL data for you. Clustering is one of the primary tasks used in the pattern recognition and data mining communities to search VL databases (including VL images) in various applications, and so, clustering algorithms that scale well to VL data are important and useful. This article compares the efficacy of three different implementations of techniques aimed at extending fuzzy c-means (FCM) clustering to VL data. Specifically, we compare methods based on (i) sampling followed by non-iterative extension; (ii) incremental techniques that make one sequential pass through subsets of the data; and (iii) kernelized versions of FCM that provide approximations based on sampling, including three proposed algorithms. it will use both loadable and VL data sets to conduct the numerical experiments that facilitate comparisons based on time and space complexity, speed, quality of approximations to batch FCM (for loadable data), and assessment of matches between partitions and ground-truth. Empirical results show that random sampling plus extension FCM, bit-reduced FCM, and approximate kernel FCM are good choices for approximating FCM for VL data. it conclude by demonstrating the VL algorithms on a data set with 5 billion objects and presenting a set of recommendations regarding the use of the different VL FCM clustering schemes.

III. 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. Note: the above equation is equivalent to

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

$$|\omega_i - X| \leq |\omega_K - X| \quad \forall k$$

Given the winning node i, the weight update is

$$\boldsymbol{\omega_k(new) = \omega_k(old) + \Delta\omega_k(n)}$$

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:

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

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.

IV. 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

1	Drink
2	Rule
3	Attention

4	Physical
5	Mistake
6	Inexperience
7	Speed
8	Road
9	Weather
10	Vehicle
11	Sight

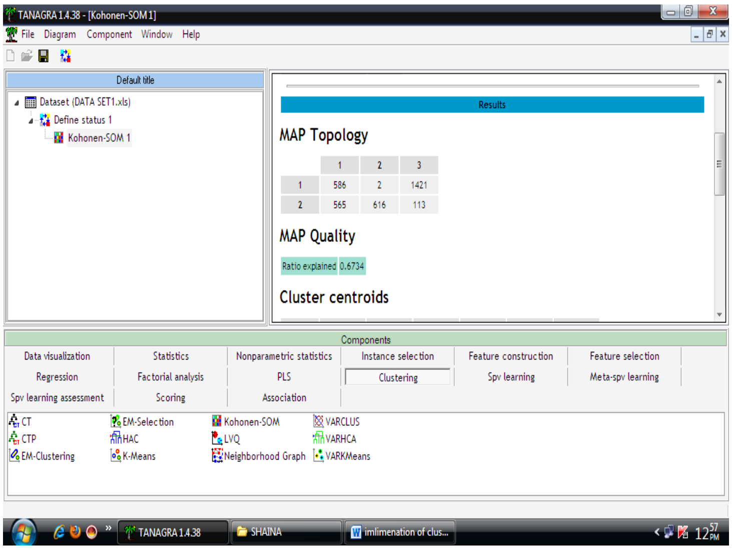

Figure 1. Implemented Kohonen SOM Algorithm and generate MAP Topology with 6 clusters, MAP Quality 0.8189

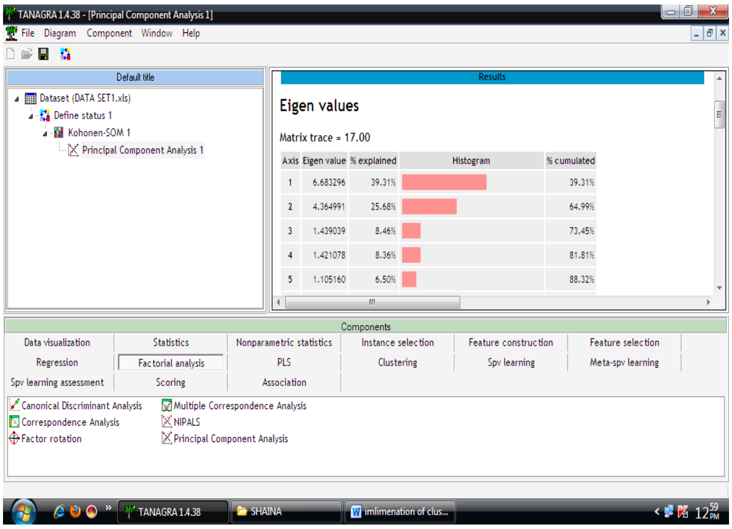

Figure 2. Implemented PCA on Kohonen SOM and evaluate Eigen Values

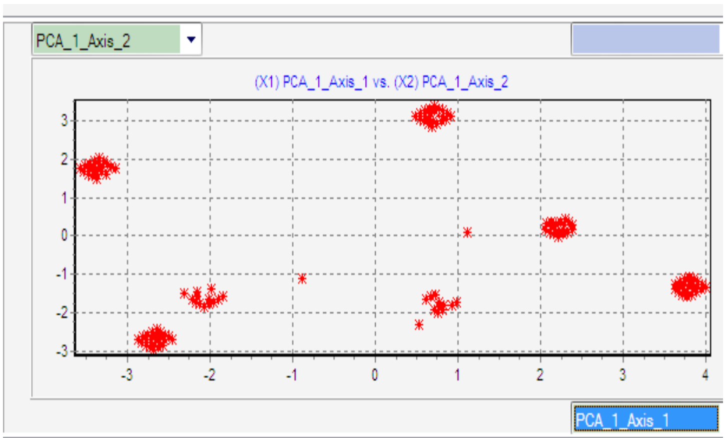

Figure 3. Now Showing 6 different clusters with same scale.

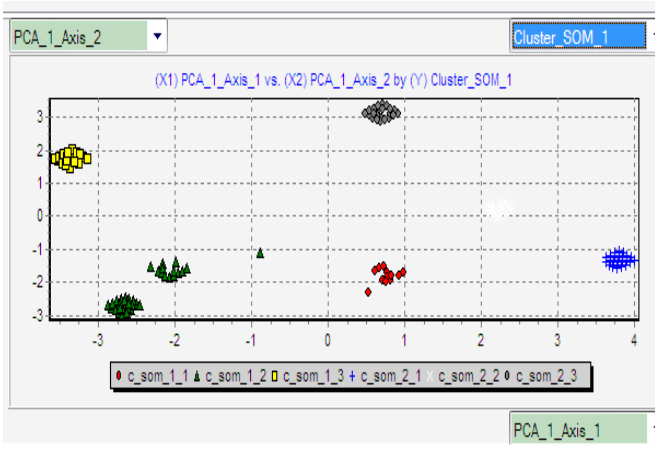

Figure 4. The KOHONEN-SOM component shows 6 different clusters with different scale

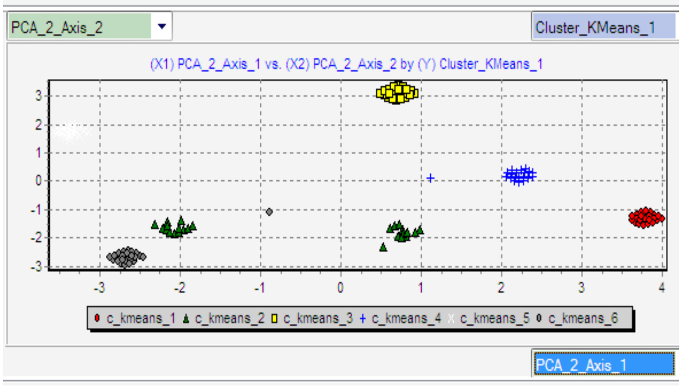

Figure 5. The K-Means algorithm shows 6 different clusters with different Conclusion

Comparison of the results between SOM and K-Means

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
No. of clusters	6	6
MAP OLOPLOGY	6	6
ERROR RATE	0.8189	0.8456
COMPUTAION TIME	297 MS	1281 MS
ACEESING TIME	FAST	SLOW

V. CONCLUSION

A series of studies have been carried out to analyze the causes of highway accidents in NH1 using data mining techniques. The objective of this project was to pre- process highway accidental data from National Highway Authority for generating human interpretable clusters and to show the advantage of using clustering approach for accidental analysis. The analysis of k-Means and SOM decision tree algorithm, whose result is compared with that obtained by the commercial program Tanagra. The analysis is conducted through three different aspects with respect to age, season and gender. The comparison shows good agreement of the results generated by the two different ways that can demonstrate the error rate and computation timey of the program.

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)

