Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Implementation and Evaluation of Improved SOM for High Dimensional Data Set

Kamalpreet Kaur Jassar*
Research Scholar, BBSBEC
Department of CSE, FGS,
Punjab, India

Dr. Kanwalvir Singh Dhindsa
Associate Professor
Department of CSE, FGS,
Punjab, India

Abstract— Clustering is a process of grouping a set of spatial objects into groups, these groups are called clusters. Clustering is a very well known technique of data mining which is mostly used method of analyzing and describing the data. It is one of the techniques to deal with the large geographical datasets. Clustering is the mostly used method of data mining. Kohonen SOM is a classical method for clustering. In this paper, a new improved approach is proposed by combining neural network and clustering algorithms. The improved Self Organizing Map algorithm, which initially starts with null network and grows with the original data space as initial weight vector, updating neighbourhood rules and learning rate dynamically in order to overcome the fixed architecture and random weight vector assignment of simple SOM. This paper illustrates performance analysis of existing SOM and improved SOM.

Keywords— Geo-referenced data, Kohonen SOM, Learning rate, Neighbourhood size, Weight vector

I. INTRODUCTION

The Kohonen SOM algorithm is a very powerful tool for data analysis [21]. SOM was originally designed to model organized connections between some biological neural networks. It was also immediately considered as a good algorithm to realize vectorial quantization, and at the same time pertinent classification, with nice properties for visualization [20]. Self-Organizing Maps (SOMs) have been used in GIScience both for clustering geo-referenced data and for the spatialization of various non-geographic datasets. The original SOM proposed by Kohonen does not take into account the particular role that geographic location has in most problems involving the clustering of geo-referenced data.

The basic idea of a SOM is to map the data patterns into n-dimensional grid of neurons or units. 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 and the output space is usually 1 or 2 D. A self-organizing map (SOM) is a kind of artificial neural network that is trained using unsupervised learning to produce a low dimensional typically two dimensional as output. It is 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 neighbourhood function to preserve the topological properties of the input space. The main set back of this technique is that the number of output nodes is predefined and only the adjacent nodes are taken as neighbourhood [8]. SOM is a clustering method because 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 other clustering algorithm such as K-Means [11]. SOM is also considered as data visualization technique because it allows to visualize data in a low dimensional representation space (basically in 2D).

In the original SOM algorithm, all variables are treated equally. When clustering geo-referenced data, spatial location is particularly important, since objects that are geographically far away should not be clustered together, even if they are similar in all other aspects. Although the term "Self-Organizing Map" could be applied to a number of different approaches (as a synonym of Kohonen's Self Organizing Map, or SOM for short, also known as Kohonen Neural Networks).

II. RELATED WORK

Spatial data mining as a knowledge discovery process has been explained by Ng and Han [2]. Thus, it plays an important role in a) extracting interesting spatial patterns and features; b) capturing intrinsic relationships between spatial and non-spatial data; c) presenting data regularity concisely and at higher conceptual levels; and d) helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance.

Sisodia et al. [3] explained about Clustering that it is an unsupervised learning task where one seeks to identify a finite set of categories termed clusters to describe the data. Various clustering algorithms of data mining have been considered and it also focuses on the clustering basics, requirement, classification problem and application area of the clustering algorithms. It also gives detail about classification of clustering techniques and their respective algorithms with the advantages and disadvantages. So this paper provides a quick review of the different clustering techniques in data mining.

Aneetha and Bose [5] proposed a modified Self Organizing Map algorithm which initially starts with null network and grows with the original data space as initial weight vector, updating neighbourhood rules and learning rate dynamically in order to overcome the fixed architecture and random weight vector assignment of simple SOM. New nodes are created using distance threshold parameter and their neighbourhood is identified using connection strength and its learning rule and the weight vector updation is carried out for neighbourhood nodes. The k-means clustering algorithm is employed for grouping similar nodes of modified SOM into k clusters using similar measures.

Halkidi et al. [16] explained the fundamental concepts of clustering while it surveys the widely known clustering algorithms in a comparative way. Moreover, it addresses an important issue of clustering process regarding the quality assessment of the clustering results. This is also related to the inherent features of the data set under concern. This is also related to the inherent features of the data set under concern. A review of clustering validity measures and approaches available in the literature is presented. This paper also illustrates the issues that are under addressed by the recent algorithms and gives the trends in clustering process.

A new approach is proposed by Berglund and Sitte [12]. The parameterless self-organizing map (PLSOM) is a new neural network algorithm based on the self-organizing map (SOM). It eliminates the need for a learning rate and annealing schemes for learning rate and neighborhood size. We discuss the relative performance of the PLSOM and the SOM and demonstrate some tasks in which the SOM fails but the PLSOM performs satisfactory.

Different data clustering algorithms has been studied and compared by Abbas [4]. These are compared according to the factors like size of dataset, type of dataset, number of clusters and tool used. The algorithms considered for investigation are k-means algorithm, self organizing map algorithm, hierarchical clustering algorithm and expectation maximization algorithm. Conclusions extracted from comparative study of these algorithms belong to the performance, quality and accuracy of algorithms.

Hosseini [19] suggests the similarities between the mechanisms used in the TASOM (Time Adaptive Self- Organizing Map) neural network and AIS (Artificial Immune Systems) are analyzed. To demonstrate the similarities, AIS mechanisms are incorporated into the TASOM network such as the weight updating is replaced by a mutation mechanism. Learning rate and neighborhood sizes are also replaced by the clonal selection process used in AIS. This new network is called TAISOM. Experimental results with TAISOM are implemented for uniform and Gaussian distributions for one and two-dimensional lattices of neurons. These experiments show that TAISOM learns its environment as expected so that neurons fill the environments quite well and the neurons also preserve the topological ordering.

III. IMPROVED SOM

In improved approach, a simple modification is done in existing SOM that completely eliminates the learning rate, the decrease of the learning rate and the decrease of the neighbourhood size [36]. A new learning rule is introduced. This has been done by making the learning rate and neighbourhood size dependant on a variable calculated from the internal state of the SOM, rather than on externally applied variables.

The learning algorithm of improved SOM is detailed below in the following steps:

Step 1 Initialize the map node's weight vectors.

Step 2 Traverse each input vector in the input data set

Step 3 Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector.

Step 4 Track the node that produces the smallest distance (this node is the best matching unit, BMU).

Step 5 Update the nodes in the neighborhood of the BMU. The new learning rule is introduced in this algorithm is as follows:

$$Wv(t+1)=Wv(t)+\theta(v,t)\alpha(t)(D(t))–Wv(t))+\theta(v,t)(t/T) \quad (1)$$

Where $\alpha(t)$ is a monotonically decreasing learning coefficient and $D(t)$ is the input vector. The neighborhood function $\theta(v,t)$ depends on the lattice distance between the BMU and neuron v. Current time is represented by t and T is the total time.

IV. PERFORMANCE ANALYSIS

A. Computational Time

The performance in terms of computational time (y-axis) for improved approach is compared with existing approach for number of iterations (x-axis) as shown in Fig. 1.

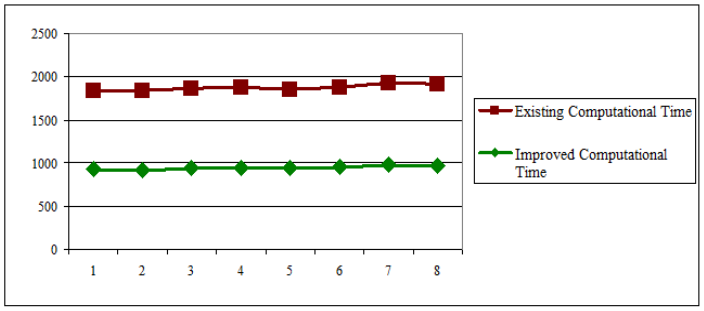

Fig.1 Computational Time Performance of Existing Vs Improved SOM

Following are the observations from comparison of computational times:

- The improved SOM consumes less computation time to fetch results than the existing SOM.
- The improved approach has minimum computation time at 5th iteration of 910 ms where in case of existing approach the minimum computation time is 918 ms obtained at 2nd iteration.
- The improved approach has maximum computation time at 8th iteration of 953 ms where in case of existing approach the maximum computational time is 976 ms obtained at 7nd iteration.
- Improved SOM adopts a new way to update weight vectors of neurons, which helps to reduce the redundancy in features extracted from the principal components. Therefore its computational time also reduces.
- The improved SOM algorithm has been modified during each training step. Therefore its adaptation is fast and the elapsed time is low, even if a large number of iterations might be necessary or the dataset is unusually large.
- The graphical representation shows the efficient results for computation time in improved SOM as compared to the existing SOM.

B. Complexity

The performance in terms of complexity time (y-axis) for improved approach is compared with existing approach for number of iterations (x-axis) as shown in Fig. 2.

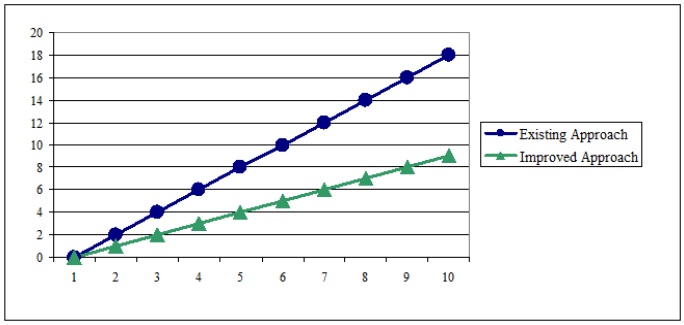

Fig. 2 Complexity Performance Measure of Existing Vs Improved SOM

Following are the observations from the graph:

- The complexity of two algorithms depends on the no. of iterations of the dataset. In the above graph the complexities are plotted against the no. of iterations.
- The improved approach has minimum complexity at 2nd iteration of 2 min where in case of existing approach the minimum complexity is 1 min obtained at 2nd iteration.
- The improved approach has maximum complexity at 10th iteration of 9 min where in case of existing approach the maximum complexity is 18 min obtained at 10th iteration.
- Complexity of improved approach is improved by 2% as compared with existing.
- The reason for less complexity is that, the improved algorithm takes less iterations steps than that of existing algorithm.
- The graphical representation shows efficient results for complexity in improved SOM as compared to the existing approach.

C. Cost

The performance in terms of complexity (y-axis) for improved approach is compared with existing approach for complexity (x-axis) as shown in Fig. 3 below.

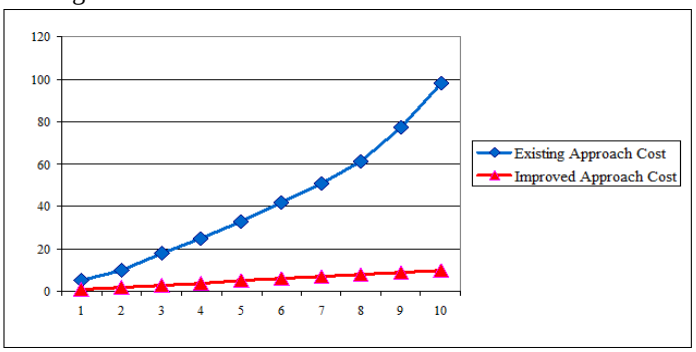

Fig. 3 Cost Incurred in Existing Vs Improved SOM

Following are the observations from the graph:

- The cost incurred for the two algorithms depends on the complexity factor and the no. of iterations in the dataset used.
- Above graph clearly shows that as the complexity of algorithm increase, the cost factor also increases.

- The improved SOM is observed in the graph to be less costly than existing SOM.
- The improved approach has minimum computation cost at 1st iteration of 1 ms whereas maximum computation cost is 10 ms obtained at 10th iteration.
- The existing approach has minimum computation time at 1st iteration of 5 ms whereas the maximum computational cost is 98 ms obtained at 10th iteration.
- Improved algorithm includes minimal additional computations per learning step, which are conveniently easy to implement and reduces the computational cost of algorithm.
- Thus the relative computational cost for the improved SOM algorithm drastically improved and cuts the learning rate almost by five times.
- The graphical representation shows efficient result of cost for improved SOM as compared to the existing approach.

D. Error Rate

The performance in terms of error rate (y-axis) for improved approach is compared with existing approach with respect to learning rate (x-axis) as shown in Fig. 4.

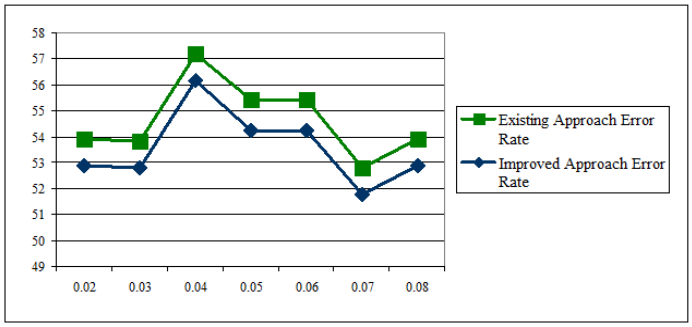

Fig. 4 Error Rate in Existing Vs Improved SOM

Following are the observations from the graph:

- The error incurred for the two algorithms depends on the learning rate factor and the no. of iterations in the dataset used.
- The improved SOM is observed in the graph to be less accurate than existing SOM.
- This is observed that, the improved approach has minimum error rate is 51.80% obtained at 0.07 learning rate. Whereas the maximum error rate is 56.15% obtained when learning rate is 0.04.
- This existing approach has minimum error rate is 52.81% obtained at 0.07 learning rate. Whereas the maximum error rate is 57.17% obtained when learning rate is 0.04.
- Compared with the existing SOM algorithm, due to adoption of such strategies as dynamic neighborhood radius, dynamic learning rate, and a new way to update weights, the improved algorithm have less error rate. Therefore it also provides a good robustness.
- The graphical representation shows efficient results for improved SOM as compared to the existing approach.

E. Efficiency

The performance in terms of efficiency (y-axis) for improved approach is compared with the existing approach for number of iterations (x-axis) as shown in Fig. 5.

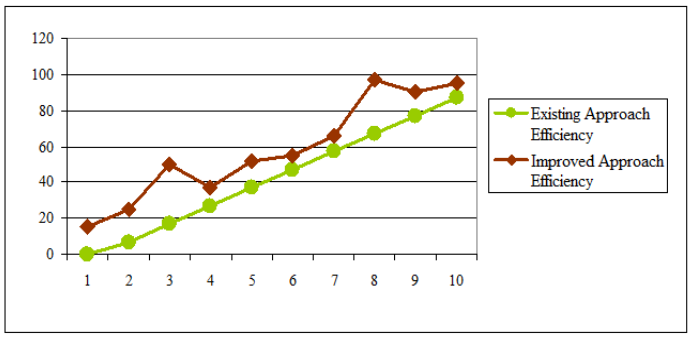

Fig. 5 Efficiency of Existing Vs Improved SOM

Following are the observations from the graph:

- The efficiency of the two algorithms depends on the complexity factor and the cost to execute the data structure.
- The efficiency of the existing approach remains constant, but it may vary in case of improved approach.
- The improved approach gives its maximum efficiency of 97% in 8th iteration. Whereas the maximum efficiency of existing approach is 87% obtained in 10th iteration.

- The improved approach has minimum efficiency of 15% obtained in 1st iteration. While the minimum efficiency of existing approach is just 2% which is also obtained in 1st iteration.
- Both algorithms possess very little difference in their efficiencies at 6th and 7th iterations.
- By reducing the number of iterations, the overall efficiency of the algorithm has been improved. It is demonstrated that the improved algorithm has the ability to reduce unnecessary computation up to 70%.
- The graphical representation shows that the improved SOM is more efficient in its performance as compared to the existing approach.

V. RESULTS

TABLE 1. COMPARATIVE RESULTS OF EXISTING AND IMPROVED SOM

Parameters	**Existing SOM**	**Improved SOM**
Computational Time	The existing approach consumes much time	The improved approach consumes half the time taken by existing approach
Complexity	The existing algorithm is more complex as it depends on learning rate	The improved algorithm is simpler to implement as it does not depends on learning rate
Cost	Existing algorithm is more expensive	Improved SOM is less expensive than existing SOM
Error Rate	It has higher error rate as compared with improved approach	It has lower error rate as compared with existing approach
Efficiency	The existing algorithm is less efficient to find clusters	The improved algorithm is more efficient than the existing approach

The results of the implementation of improved SOM algorithm has led to some important conclusions. This approach relies on the idea that the learning rate and neighbourhood size should not vary according to the iteration number, but rather vary according to how well the map represents the topology of the input space. It also markedly decreases the number of iterations required to get a stable and ordered map. Improved SOM completely eliminates the selection of the learning rate, the annealing rate and annealing scheme of the learning rate and the neighbourhood size, which have been an inconvenience in applying SOMs. It learns continuously from its environment, and only a one-time initialization is needed to work in its changing environment. The improved SOM also reduces the training time and preserves generality.This is achieved without inducing a significant computation time or memory overhead. It also covers a greater area of the input space, leaving a smaller gap along with the edges.

VI. CONCLUSIONS

The brief conclusion of analysis is as follows:

- The improved SOM approach relies on the idea that the learning rate and neighbourhood size should not vary according to the iteration number, but rather vary according to how well the map represents the topology of the input space.
- It also markedly decreases the number of iterations required to get a stable and ordered map.
- Improved SOM completely eliminates the selection of the learning rate, the annealing rate and annealing scheme of the learning rate and the neighbourhood size, which have been an inconvenience in applying SOMs.
- It learns continuously from its environment, and only a one-time initialization is needed to work in its possibly changing environment.
- The improved SOM also reduces the training time and preserves generality. This is achieved without inducing a significant computation time or memory overhead.
  It also covers a greater area of the input space, leaving a smaller gap along the edges.

VII. FUTURE SCOPE

The improved SOM can be applied to many familiar problems. The future scope of this is listed below:

- Improved SOM can be used to determine the direction of the sound source and orienting the microphones toward the sound source.
- It can deal with cases where the number of input dimensions is far higher than the number of output dimensions.
- The SOM related methods are finding wide application in more and more fields to make the methods more efficient, consistent and robust, especially for large-scale and real-world applications.
- Improved SOM can be used in image segmentation because it can handle high-dimensional and clustered data.
- For general pattern recognition, it may have more potential than implied by current practice, which often limits the SOM to a 2-D map and empirically chosen model parameters.
- Extensions of the SOM may provide useful tools in deciphering and interpreting the information content and relationships conveyed among stimuli and responses.

- The algorithm can be further extended in order to deal with complex biological signals and networks. For example in handling spike trains.

REFERENCES

[1] Agrawal, R., Mehta, M., Shafer, J., Srikant, R., Arning, A. and Bollinger, T, "The Quest Data Mining System", Proceedings of 1996 International Conference on Data Mining and Knowledge Discovery (KDD'96), Portland, Oregon, pp. 244-249, 1996.

[2] Ng, R.T. and Han, J., "Clarans: A Method For Clustering Objects For Spatial Data Mining", IEEE Transactions On Knowledge And Data Engineering, Vol.14, No.5, pp. 1003-1016, 2002.

[3] Sisodia, D., Singh, L., Sisodia, S. and Saxena, K., "Clustering Techniques: A Brief Survey of Different Clustering Algorithms", International Journal of Latest Trends in Engineering and Technology, Vol. 1, Issue 3, pp. 82-87, 2012.

[4] Abbas, O.A., "Comparison between Data Clustering Algorithms", The International Arab Journal of Information Technology, Vol. 5, No. 3, pp. 320-325, 2008.

[5] Aneetha, A.S. and Bose, S., "The combined approach for anomaly detection using neural networks and clustering techniques", Computer Science & Engineering: An International Journal, Vol. 2, No. 4, pp. 37-46, 2012.

[6] Sharma, H. and Kaler, N.K., "A Synthesized Approach for Comparison and Enhancement of Clustering Algorithms in Data Mining for Improving Feature Quality", International Journal of Soft Computing and Engineering, Vol. 4, Issue 2, pp. 114-117, 2014.

[7] Toor, A.K. and Singh, A., "Analysis of Clustering Algorithms Based on Number of Clusters, Error Rate, Computation Time and Map Topology on Large Data Set", International Journal of Emerging Trends & Technology in Computer Science, Vol. 2, Issue 6, pp. 94- 98, 2013.

[8] Bacao, F., Lobo, V. and Painho, M., "Self-organizing Maps as Substitutes for K-Means Clustering", International Conference on Computational Science, Springer-Verlag Berlin Heidelberg, Vol. 3516, pp. 476 – 483, 2005.

[9] Ravikumar, S. and Shanmugam, A., "Comparison of SOM Algorithm and K- Means Clustering Algorithm in Image Segmentation", International Journal of Computer Applications, Vol. 46, No. 22, pp. 21-25, 2012.

[10] Sharma, K. and Dhiman, R., "Implementation and Evaluation of K-Means, Kohonen-SOM, and HAC Data Mining Algorithms base on Clustering", International Journal of Computer Science Engineering & Information Technology Research, Vol. 3, Issue 1, 2013, pp. 165-174.

[11] Bhatia, S.K. and Dixit, V.S., "A Propound Method for the Improvement of Cluster Quality", International Journal of Computer Science Issues, Vol. 9, Issue 4, No. 2, pp. 216-222, 2012.

[12] Berglung, E. and Sitte, J., "The Parameter-less Self-Organizing Map Algorithm", IEEE Transactions on Neural Network, Vol. 17, No. 2, pp. 305-316, 2006.

[13] Chen, Y., Qin. B., Liu, T., Liu, Y. and Li, S., "The Comparison of SOM and K- means for Text Clustering", Computer and Information Science, Vol. 3, No. 2, pp. 268-274, 2010.

[14] Dhingra, S., Gilhotra, R. and Ravishanker, R., "Comparative Analysis of Kohonen-SOM and K-Means data mining algorithms based on Academic Activities", International Journal of Computers & Technology, Vol. 6, No. 1, pp. 237-241, 2013.

[15] Ganda, R. and Chahar, V., "A Comparative Study on Feature Selection Using Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 9, pp. 26-33, 2013.

[16] Halkidi, M., Batistakis, Y. and Vazirgiannis, M., "On Clustering Validation Techniques", Journal of Intelligent Information Systems, Vol.17, No. 2, pp. 107–145, 2001.

[17] Raghuwanshi, S.S. and Arya, P.N., "Comparison of K-means and Modified K-means for Large Data-set", International Journal of Computing, Communications and Networking, Vol. 1, No. 3, pp. 106-110, 2012.

[18] Sumathi, N., Geetha, R. and Bama, S.S., "Spatial data mining-techniques trends and its applications", Journal of Computer Applications, Vol.1, No.4, pp. 28-30, 2008.

[19] Hosseini, H.S., "The Time Adaptive Self-Organizing Map is a Neural Network Based on Artificial Immune System", Proceedings of IEEE World Congress on Computational Intelligence, International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, Canada, Vol. 6, No. 3, pp.1007-114, 2006.

[20] Hemalatha, M. and Saranya, N.N., "A Recent Survey on Knowledge Discovery in Spatial Data Mining", International Journal of Computer Science Issues, Vol. 8, Issue 3, pp. 473-479, 2011.

[21] Johal, H.S., Singh, B., Singh, H., Nagpal, A. and Virdi, H.S.,"Using Kohonen-SOM & K-Means Clustering Techniques to Analyze QoS Parameters of RSVP", Proceedings of the World Congress on Engineering and Computer Science, Vol. 1, pp. 431-436, 2012.

[22] Kohonen, T., "The self-organizing maps", Proceedings of the IEEE, Vol. 78, No. 9, pp. 1464-1480, 1990.

[23] Kaur, J. and Singh, G., "Review of Error Rate and Computation Time of Clustering Algorithms on Social Networking Sites", International Journal of Computer Application, Vol. 113, No. 8, pp. 32-35, 2015.

[24] Mingoti, S.A. and Lima, J.O., "Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms", European Journal of Operational Research, Vol. 174, pp. 1742–1759, 2006.

[25] Sundararajan, S. and Karthikeyan, S., "A Study On Spatial Data Clustering Algorithms In Data Mining", International Journal Of Engineering And Computer Science, Vol. 1, Issue 1, pp. 37-41, 2012.

[26] Subitha, N. and Padmapriya, A., "Clustering Algorithm for Spatial Data Mining: An Overview", International Journal of Computer Applications, Vol. 68, No.10, pp. 28-33, 2013.

[27] Murugavel, P. and Punithavalli, M., "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", International Journal on Computer Science and Engineering, Vol. 3, No. 1, pp. 333-339, 2011.

[28] Bacao, F., Lobo, V. and Painho, M., "Clustering census data: comparing the performance of self-organizing maps and k-means algorithms", Proceedings of KDNet (European Knowledge Discovery Network of Excellence), Knowledge-Based Services for the Public Sector, Workshop 2: Mining Official Data, Petersberg Congress Hotel, Bonn, Germany, Vol. 2, 2004.

[29] Kohonen, T., "Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics, Springer, Espoo, Finland, Vol. 43, pp. 59-69, 1982.

[30] Kohonen, T., "Essentials of the self-organizing maps", International Conference on Neural Networks, Vol. 37, pp. 52-65, 2013.

[31] Birdi, M., Gangwar, R.C. and Singh, G., "A Data Mining Clustering Approach for Traffic Accident Analysis of National Highway-1", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 10, pp. 44- 47, 2014.

[32] Mishra, M. and Behera, H.S., "Kohonen Self Organizing Map with Modified K-means Clustering for High Dimensional Data Set", International Journal of Applied Information Systems, Vol. 2, No. 3, pp. 34-39, 2012.

