Efficient Data Retrieval using Combine Approach of SOM and K-Mean Clustering

Deepa Sharma
Department of Computer Science & engineering
Samrat Ashok Technological Institute
Vidisha, MP, India

ABSTRACT
Emergence of recent techniques for scientific knowledge collection has resulted in large scale accumulation of information relating various fields. Typical info querying ways are inadequate to extract helpful data from huge knowledge banks. Cluster analysis is one of the key knowledge analysis way and the k-means clustering algorithm is widely used for several data mining applications. The analysis of the cancer data set with the k mean and then applying with the Som. Many ways are planned within the literature for improving the performance with the k-means clustering formula. This paper proposes a technique for creating knowledge retrieval more practical and efficient using som with K mean clustering technique, So as to get better clustering with reduced quality**.**

Keywords
Clustering, Health Care, SOM, K-Mean, False detection.

1. INTRODUCTION
Advances in scientific knowledge assortment strategies have resulted within the massive scale accumulation of promising knowledge relating numerous fields of science and technology. Because of the development of novel techniques for generating and assembling knowledge, the speed of growth of scientific databases has become tremendous. Therefore it is not possible to extract helpful data from them by exploitation standard information analysis techniques. Effective mining strategies are fully essential to search out implicit data from vast databases. The k-means formula [1, 2, 3, 4, 5] is effective in manufacturing clusters for several sensible applications. However the procedure quality of the initial k-means formula is extremely high, particularly for big knowledge sets. Moreover, this formula leads to totally different variety of clusters looking on the random selection of initial centroids. Various attempts were made by researchers for improving the performance of the k-means clustering algorithm. This paper deals with a method for improving the accuracy and efficiency of the k-means algorithm.

One method of social media data processing is link (relationship) mining, which combines social networks, link analysis, hypertext and internet mining, graph mining, relative learning, and inductive logic programming [6]. Researching links involve many steps: link-based object classification (categorizes objects primarily based on links and attributes) [7], object kind prediction (predicts object types supported attributes, links, and objects joined to it) [8], link kind prediction (predicts the aim of the link supported the objects involved) [9], link existence prediction (predicts the existence of a link) [10], link cardinality estimation (predicting the amount of links (and objects reached) to an object) [11], object reconciliation (determining whether or not 2 objects are the same supported their links) [12], cluster detection (predicting if an object set belongs together) [13], sub-graph detection (discovering sub-graphs inside networks) [14], and data mining (mining for information concerning data) [15], [16]. Other samples of mining social networks are link prediction, namely exploitation the options intrinsic of the present model of a social network to model future connections inside the network. Viral promoting uses spoken impact by measurement the interactions among customers and thoroughly promoting people with the foremost social connections. Newsgroups discussions take advantage of "response" relationships supported however typically people answer messages they agree (or disagree) with exploitation graph-partitioning algorithms. Relation choice and extraction of a multi-relational network measures and ranks completely different relations based on user data (acquired through queries). Traditional social sciences use surveys and involve subjects in the information assortment method restricted by this method, information collected are of tiny sizes and generally many subjects in one study. In contrast, thousands of users of social media manufacture inordinate amounts of information with wealthy user interactions.

Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to simplify their further processing. Data clustering has many engineering applications including the recognition of part families for cellular manufacture. Clustering is the process of classifying data objects into a set of disjoint classes called clusters. Clustering is an example of unsupervised classification. Classification refers to a procedure that partition data objects to a set of classes. Unsupervised means that clustering does not depends on predefined classes and training examples while classifying the data objects. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [17], [18]. Therefore, a cluster is a assortment of objects that are similar among themselves and dissimilar to the objects belonging to other clusters. Clustering is an crucial area of research, which finds applications in many areas including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc. Cluster analysis is a one of the key data analysis tool in the data mining. Clustering algorithms are mainly comes into two categories: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm splits the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step [18].

2. RELATED WORK
Recently much research has been done in the field of efficient healthcare data mining. Prominent research work in this field by different researchers is as given below:

Fang Yuan et al. [22] proposed a systematic method for finding the initial centroids. The centroids obtained by this method are consistent with the distribution of data. Hence it produced clusters with better accuracy, compared to the original k-means algorithm. However, Yuan's method does not suggest any improvement to the time complexity of the k-means algorithm.

Fahim A M et al. [23] proposed an efficient method for assigning data-points to clusters. The original k-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose-one similar to the k-means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original k-means algorithm. Hence there is no guarantee for the accuracy of the final clusters.

Identifying cancer risks [24] related to medicative agents plays a crucial role in cancer management and hindrance. Case reports of cancers related to pharmacotherapy are escalating within the Food and Drug Administration Adverse Event coverage System (FAERS). the target of this study is to assess the chance of pancreatic cancer related to anti-diabetic medication of dipeptidyl peptidase four (DPP 4) inhibitors with or while not combination of antidiabetic drug. Using the FAERS public information, the adverse event reports (ADRs) related to wide used DPP 4 inhibitors with or while not combination of Glucophage were generated and evaluated. Standardized pharmacovigilance tools were applied to find the signal for cancer risks by calculative the proportional reporting ratio (PRR) and also the reporting odds ratio (ROR). Among 12618 ADRs related to sitagliptin from 2007 to 2011, there have been 223 cases of cancer. There was a big correlation between the cancer coverage magnitude relation and also the time (R=0.796, P<0.001). Pancreatic cancers accounted for twenty second of all combined cancer adverse events. Pharmacovigilance assessment from 2007 to 2012 indicated that there was a big risk of carcinoma related to DPP four inhibitors treatment (ROR=5.922). Curiously, negligible risk of carcinoma risk was related to Glucophage (ROR=1.214). Combination of DPP four matter sitagliptin with Glucophage correlates with considerably lower risk of carcinoma compared to sitagliptin treatment while not Glucophage (OR=0.277, 95%CI: 0.210-0.365).

The Self-Organizing Map (SOM) [25] toolbox is a powerful toolbox. The vector quantization technique that places image vectors on a regular low dimensional grid makes the SOM a robust image tool . The SOM toolbox is an implementation of the SOM and its visualization within the Matlab five computing surroundings. In this article, the SOM toolbox and the various functions which the SOM toolbox contains are shortly presented. The performance of SOM Toolbox in terms of machine load is evaluated and compared to corresponding c program. The SOM could be a very good tool to visualize high dimensional knowledge. It's most fitted for data understanding section of the data discovery methodology, though it is often used for data preparation, modeling and classification furthermore. The analysis considers the quantitative analysis of SOM mappings, especially analysis of clusters and their properties. New functions and graphical program tools are going to be further to the Toolbox to extend its quality in data processing. Also outside contributions to the toolbox are welcome. The SOM toolbox promotes the utilization of SOM formula – in analysis furthermore as in industry – by creating its best options additional promptly accessible.

Lung cancers [26] caused by activating mutations within the epidermic growth factor receptor (EGFR) square measure at the start alert to little molecule tyrosine kinase inhibitors (TKIs), however the effectualness of those agents is commonly restricted due to the emergence of drug resistance conferred by a second mutation, T790M. threonine 790 is that the ''gatekeeper'' residue, a vital determinant of matter specificity within the ATP binding pocket. The T790M mutation has been thought to cause resistance by sterically block binding of TKIs like gefitinib and erlotinib, however this rationalization is tough to reconcile with the very fact that it remains sensitive to structurally similar irreversible inhibitors. Here, They tend to show by employing a direct binding assay that T790M mutants retain low-nanomolar affinity for gefitinib. moreover, They tend to show that the T790M mutation activates WT EGFR which introduction of the T790M mutation increases the ATP affinity of the oncogenic L858R mutant by additional than an order of magnitude. The increased ATP affinity is that the primary mechanism by that the T790M mutation confers drug resistance. Crystallographic analysis of the T790M mutant shows how it will adapt to accommodate tight binding of various inhibitors, including the irreversible matter HKI-272, and conjointly suggests a structural mechanism for chemical action activation. They tend to conclude that the T790M mutation could be a ''generic'' resistance mutation that may reduce the efficiency of any ATP-competitive enzyme matter and that irreversible inhibitors overcome this resistance merely through covalent binding, not as a results of an alternate binding mode.

Katherine Faust et al. [27] presented limitations in utility of the triad census for studying similarities among native structural properties of social networks. A triad census compactly summarizes the native structure of a network using the frequencies of sixteen isomorphy categories of triads (sub-graphs of three nodes). The empirical base for this study may be a assortment of fifty one social networks measuring different relative contents (friendship, advice, agonistic encounters, victories in fights, dominance relations, and so on) among a range of species (humans, chimpanzees, hyenas, monkeys, ponies, cows, and variety of bird species). Results show that, in combination, similarities among triad censuses of those empirical networks are for the most part explained by nodal and two properties – the density of the network and distributions of mutual, asymmetric, and null dyads. These results prompt us that the diversity of network-level properties is very forced by the dimensions and density of the network and caution ought to be taken in interpreting higher order structural properties after they are for the most part explained by native network options.

A novel data processing methodology [28] was developed to determine the expertise of the drug Sitagliptin (trade name Januvia) by diabetes patients. They have a tendency to devised a two-step analysis framework. Initial explorative analysis was performed on som to work out structures based mostly on user opinions among the forum posts. The results were assortment of user's clusters and their correlative (positive or negative) opinion of the drug. Consequent modeling victimization network analysis methods confirm powerful users among the forum members. These findings will offer new direction of analysis into rapid information assortment, feedback, and analysis which will alter improved outcomes and solutions for public health and vital feedback for the manufacturer. The goal of this study was to rework the posts of a forum dedicated to diabetes patients into vectors to be ready to intelligently mine user opinion of the drug Sitagliptin. The results open new options, and demands, into developing additional comprehensive solutions throughout this space.

3. K-MEAN CLUSTERING ALGORITHM & SOM

This section describes the k-mean clustering algorithm. K-mean clustering algorithm classify the given dataset into k-number of disjoint clusters. The value of k is fixed and known in advance. The algorithm be composed of two separate section. In the first section the number of k centroid is fixed and in the next section all the point of dataset are assigned to nearest centroid. The

distance between the data point and the centroid is calculated by Euclidean distance.

When all the points are addded in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the addition of new points may lead to modification in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, making a loop. As a result of this loop, the k centroids may adjust their position in a step by step manner. finally, a situation will be reached where the centroids do not move anymore. This denotes the convergence criterion for clustering. The k-means algorithm is the most extensively studied clustering algorithm and is practical in producing good results. The significant constraint of this algorithmn is that it make distinctive clusters for various sets of values of the initial centroids. Nature of the final bunches relies on upon the selection of the initial centroids. The k-implies algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of iterations and the number of clusters.

SOMs are neural networks that presents low-dimensional representation of high-dimensional data [19]. SOM training is based on unsupervised learning algorithm. This network is made up of two layers of units one layer is input unit and another layer represents output unit. Input units are connected to output units with weights. The weight values reflect on the cluster content. The SOM displays the data to the network, draw together similar data weights to similar neurons. The benefits and capabilities have been demonstrated where in spite of the reduction of the space size, the information, and identification schema of the clusters remained the same [20]. When new data is fed into the network, the closest weights matching the data change to reflect the new data. The neurons farther from the new data rarely modify. This process continues until data is no longer fed, resulting in a two-dimensional map. The SOM toolbox (www.cis.hut.fi/projects/somtoolbox) [21] is used. The SOM was trained using various map sizes, using quantization and topographic errors as validation measures. The former is the result of the average distance between every input vector and its best matching neuron (BMN), in addition to evaluating how the trained map fits into the input data [19]. The latter uses the structure of the map to conserve its topology by representing its accuracy: it is calculated using the proportion of the weights for the first and second BMNs are farther than required for evaluating the topology.

4. PROPOSED ARCHITECTURE

The process starts with loading the cancer dataset into the SOM and generating the weight vector graph. The dataset is then filtered for the empty values. The empty value set are then removed and the each expression returns the truth value. The absolute expression is then identified form the dataset.

The figure I. shows the proper working architecture of the proposed method. This architecture shows the complete working process of the proposed methodology.. The filtered dataset is then generated and passed to the K- mean algorithm. The modified dataset is then used for the data retrieval. The analysis of false detection and redundancy analysis is done with the original dataset and the modified dataset. The modified dataset is also passed to the SOM. The analysis of time complexity and accuracy analysis is done with the self organizing map.

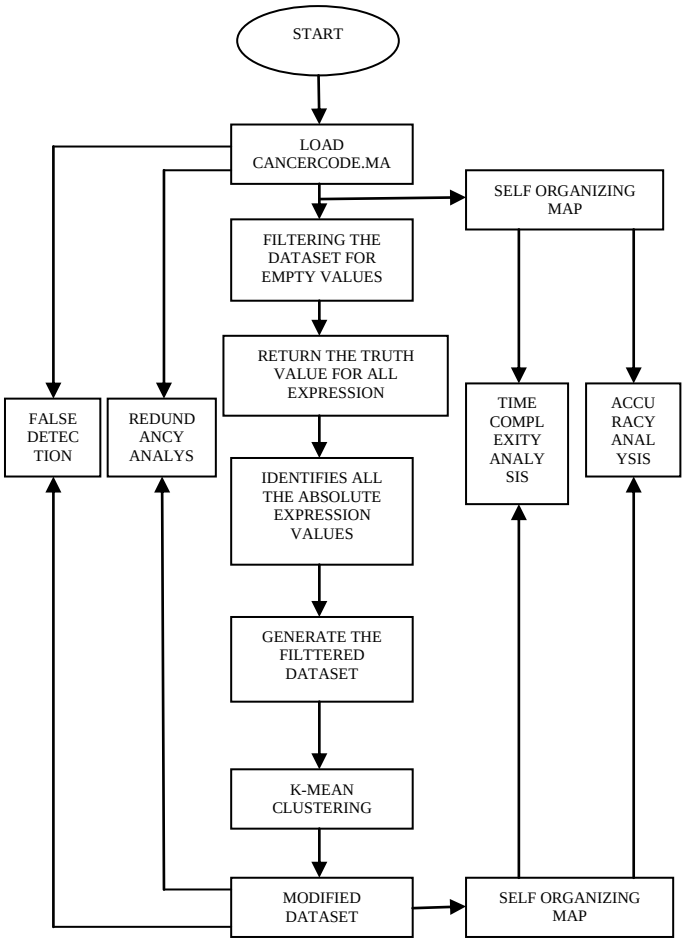

Figure I. Working Architecture

5. PROPOSED PSEUDO CODE

Step 1: Load the cancer dataset into the SOM.

Step 2: Generate the weight vector for the SOM.

```
plotsom(net.iw{1,1},net.layers{1}.distances);
```

Step 3: Filter the Dataset for empty values by strcmp.

```
emptySpots = strcmp('EMPTY',tdata);
```

Step 4: Use the |genevarfilter| function to filter out genes with small variance over time.

Step 5: Use |geneentropyfilter| to remove genes whose profiles have low entropy.

Step 6: clusters = cluster(clusterTree, 'maxclust', 16);

Step 7: kmeans(data, 16, 'dist','corr', 'rep',5,... 'disp','final');

Step 8: result1 = pcvars./sum(pcvars) * 100

```
result = cumsum(pcvars./sum(pcvars) * 10)
plotsom(net.iw{1,1},net.layers{1}.distances);
r1 = clusterTree (:,3)* 100
r3 = sort(cidx)
r2 = sort (clusters)
```

Step 9: Plot the result analysis graph.

Figure II. Pseudo Code for the Proposed Methodology

6. RESULT ANALYSIS

In the proposed work the dataset is first filtered for the null values and then it is merged with K-Mean algorithm and then the dataset is loaded to self organized map to train the network. The following are the result analysIs which are being generated and compared with old data set.

In this graph shows the false detection analysis, false positive or false alarm is indicate the given condition fulfilled for data retrieve, but actually data retrieval condition are not fulfilled. In this graph shows false detection at the time of SOM and K-mean with SOM case where x-axis shows data set and y-axis percentage of false detection. At the time of SOM percentage of false data retrieval is greater than the proposed approach k-mean with SOM because SOM take the input data set as raw data or huge data so possibility of error is more, graph is proportional shows maximum error while data set increases.

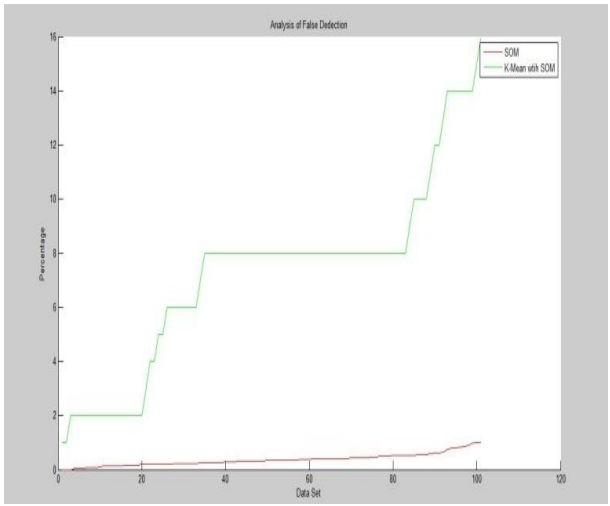

Figure III. False Detection

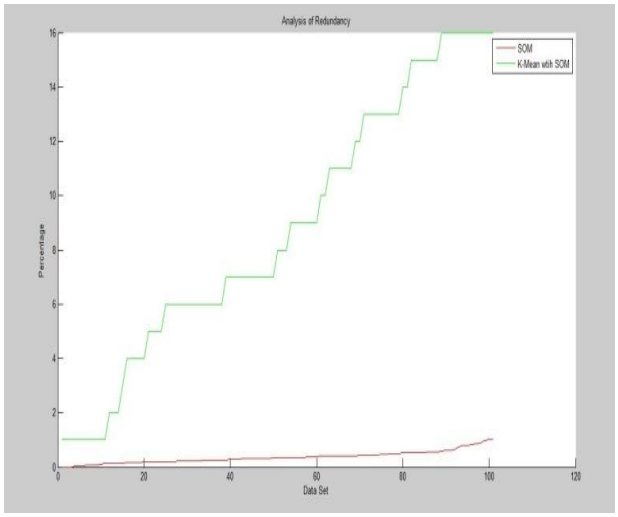

Figure IV. Redundancy Analysis

Redundancy Analysis is non symmetric method which allows studying the relationship between variables Y and X . While the Canonical Correlation Analysis is a symmetric method. In Canonical Correlation Analysis, the components extracted from both variable X and Y are such that their correlation is maximized. In Redundancy Analysis, the components extracted from X are such that they are as much as possible correlated with the variables of Y. Then, the components of Y are extracted so that they are as much as possible correlated with the components extracted from X. The X axis shows the data set and the Y axis shows the Percentage. The green is for the K-Mean with the SOM and the red line implies the direct implementation of SOM. The result shows that K-mean decrease the redundancy in the dataset. The accuracy analysis shows that the with respect to time the correct information retrieval form the dataset.

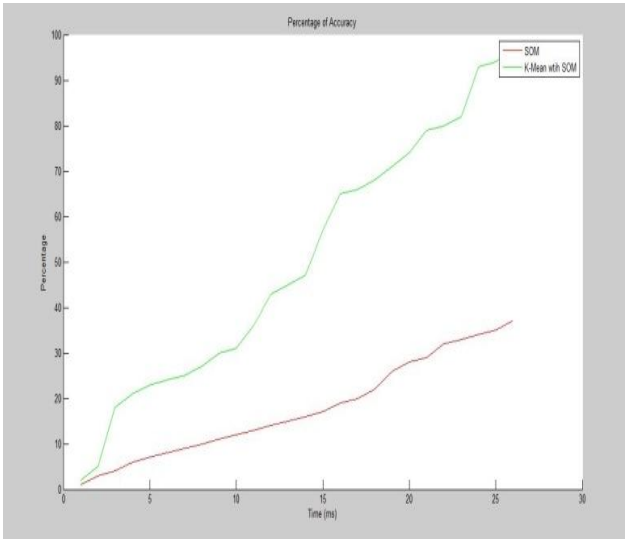

Figure V. Percentage of Accuracy

The graph here represents the percentage of the accuracy comparing the SOM and K-Mean with SOM. The green line represents the K-Mean with the SOM and the red line represents the SOM on normal dataset. The graph shows that the K-Mean with SOM has high accuracy as compared to the normal SOM because the K-Mean with SOM data set has less redundancy and filtered data set.

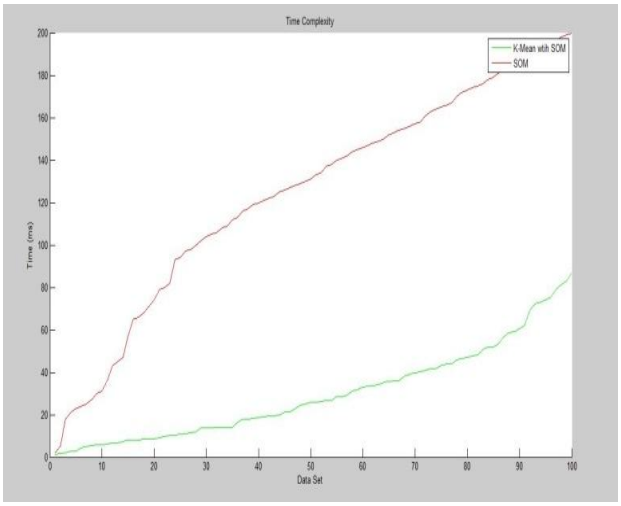

Figure VI. Time Complexity

In the above graph analyze time complexity or processing time in mille seconds of algorithm just in case of self organizing maps (existing) and k-mean with self organizing maps planned case. This graph conclude that our planned methodology take longer for retrieval inferences from data set as a result of its uses two layer of process particularly pre-processing for self organizing maps implementation at that time pass to k-mean agglomeration for fine grain retrieval with correct manner. Figure VII show the weight vector at the time of K-mean with SOM (Proposed) case, during n^{th} iteration and get maximum time weight vector are

nearly same it means network training is perfectly with less error rate.

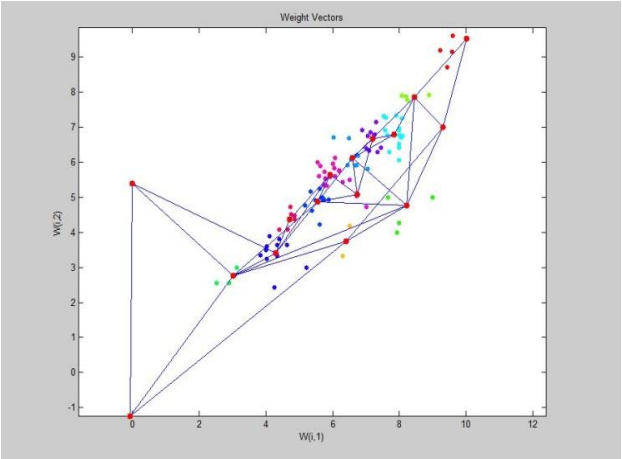

Figure VII. Weight Vector for K-Mean with SOM

While network weight vector is more change or fluctuate, it means number of epoch or trial of training is more complex for accurate data retrieval.

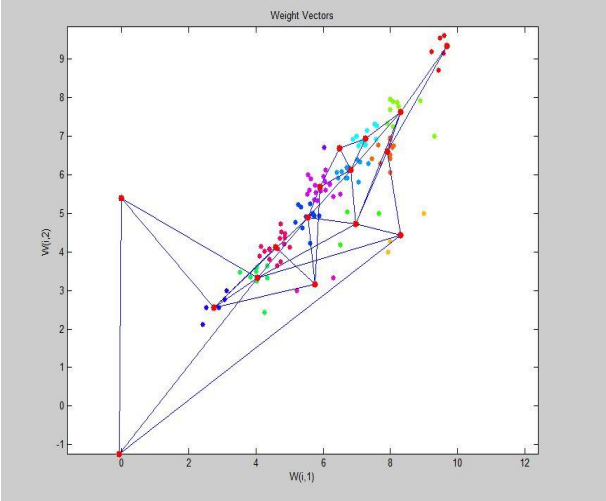

Figure VIII. Weight Vector with SOM

In the figure VIII also show the weight vector changes in every epoch while network at the time of training of SOM, in this graph weight vector are maximum change in every time that means our network training time is grater as well as un-stable. From the weight vector we conclude that our processing time while network training stage is grater and accuracy of network is very poor as compare of proposed approach.

7. CONCLUSION

The result analysis shows that the performance of the SOM improves with the application of K-mean clustering algorithm as compared to normal data set. The improvement in the weight vectors is clearly implies that the SOM with K means performs better as compared to SOM implied to normal data set. Specific algorithms will perform network-clustering, one amongst the elemental tasks in network analysis. In future we will try to train the SOM for the decision making so as to retrieve the best possible result to the future problem. So that the Physicians may collect necessary feedback (stored within the labels characterizing these network modules) from different doctors and patients that may facilitate them in their treatment recommendations and thereby up the treatment results and finally patients might value and leverage different consumers data before creating better-informed aid decisions.

8. REFERENCES

[1] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

[2] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.

[3] McQueen J, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., (1):281–297, 1967.

[4] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.

[5] Stuart P. Lloyd, "Least squares quantization in pcm," IEEE Transactions on Information Theory, 28(2): 129-136.

[6] L. Getoor and C. Diehl, "Link mining: A survey," SIGKDD Explor. Newslett., vol. 7, pp. 3–12, Dec. 2005.

[7] Q. Lu and L. Getoor, "Link-based Classification," in Proc. 20th Int. Conf. Machine Learning, Washington, DC, USA, 2003, pp. 496–503.

[8] A. Ng, A. Zheng, and M. Jordan, "Stable algorithms for link analysis," in Proc. SIGIR Conf. Inform. Retrieval, New Orleans, LA, USA, 2001, pp. 258–266.

[9] B. Taskar, M. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," presented at the Adv. Neural Inform. Process. Syst., Vancouver, Canada, 2003.

[10] D. Liben-Nowell and J. M. Kleinberg, "The link prediction problem for social networks," J. Amer. Soc. Inform. Sci. Technol., vol. 57, pp. 556–559, May 2007.

[11] Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid, "Links and paths through life sciences data sources," in Proc. 1st Int.Workshop Data Integr. Life Sci., Leipzig, Germany, 2004, pp. 203–211.

[12] J. Noessner, M. Niepert, C. Meilicke, and H. Stuckenschmidt, "Leveraging terminological structure for object reconciliation," in The Semantic Web: Research and Applications. Berlin, Germany: Springer, 2010, pp. 334–348.

[13] M. E. J. Newman, "Detecting community structure in networks," Eur. Phys. J., vol. 38, pp. 321–330, Mar. 2004.

[14] J. Huan and J. Prins, "Efficient mining of frequent subgraphs in the presence of isomorphism," in Proc. 3rd IEEE Int. Conf. Data Mining, Melbourne, FL, USA, 2003, pp. 549–552.

[15] D. Hand, "Principles of data mining," Drug Safety, vol. 30, pp. 621–622, Jul. 2007.

[16] J. Hans and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2006.

[17] S. Deelers and S. Auwatanamongkol, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," International Journal of Computer Science, Vol. 2, Number 4.

[18] Margaret H Dunham, Data Mining-Introductory and Advanced Concepts, Pearson Education, 2006.

[19] I. Mierswa, M. Wurst, W. Michael, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Philadelphia, PA, USA, 2006, pp. 935–940.

[20] P. Bonato, P. J. Mork, D. M. Sherill, and R. H. Westgaard, "Data mining of motor patterns recorded with wearable technology," IEEE Eng. Med. Biol. Mag., vol. 22, no. 3, pp. 110–119, May/Jun. 2003.

[21] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Self- Organizing Map in MATLAB: The SOM Toolbox," in Proc. Matlab DSP Conf., Espoo, Finland, 1999, pp. 35–40.

[22] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.

[23] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):1626–1633, 2006.

[24] Xiaodong Feng1, Amie Cai, Kevin Dong, Wendy Chaing, Max Feng, Nilesh S Bhutada, John Inciardi, Tibebe Woldemariam Feng, "Assessing Pancreatic Cancer Risk Associated with Dipeptidyl Peptidase 4 Inhibitors: Data Mining of FDA Adverse Event Reporting System (FAERS)", J Pharmacovigilance 2013, http://dx.doi.org/10.4172/2329-6887.1000110.

[25] Juha Vesanto, Johan Himberg, Esa Alhoniemi and Juha Parhankangas, "Self-organizing map in Matlab: The SOM Toolbox", Proceedings of the Matlab DSP Conference 1999, Espoo, Finland, November 16–17, pp. 35–40, 1999.

[26] Cai-Hong Yun, Kristen E. Mengwasser, Angela V. Toms, Michele S. Woo, Heidi Greulich, Kwok-Kin Wong, Matthew Meyerson, Michael J. Eck, "The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP", Pp. 2070–2075, PNAS, February 12, 2008, vol. 105 no.6,www.pnas.org_cgi_doi_10.1073_pnas.0709662105.

[27] Katherine Faust, Metodoloski Zvezki, "Comparing Social Networks: Size, Density, and Local Structure", Vol. 3, No. 2, 2006, 185-216.

[28] Altug Akay, Andrei Dragomir, Bjorn Erik Erlandsson, "A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin", IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 1, Pp. 2168-2194, January 2015.