Efficient Genetic K-Means Clustering for Health Care Knowledge Discovery

Ahmed Alsayat
Department of Computer Science
Bowie State University
asayat@ju.edu.sa

Hoda El-Sayed
Department of Computer Science
Bowie State University
helsayed@bowiestate.edu

Abstract—**Data mining and machine learning are becoming the most interesting research areas and increasingly popular in health organizations. The hidden patterns among patients data can be extracted by applying data mining. The techniques and tools of data mining are very helpful as they provide health care professionals with significant knowledge toward a decision. Researchers have shown several utilities of data mining techniques such as clustering, classification, and regression in health care domain. Particularly, clustering algorithms which help researchers discover new insights by segmenting patients and providing them with effective treatments. This paper, reviews existing methods of clustering and present an efficient K-Means clustering algorithm which uses Self Organizing Map (SOM) method to overcome the problem of finding number of centroids in traditional K-Means. The SOM based clustering is very efficient due to its unsupervised learning and topology preserving properties. Two-staged clustering algorithm uses SOM to produce the prototypes in the first stage and then use those prototypes to create clusters in the second stage. Two health care datasets are used in the proposed experiments and a cluster accuracy metric was applied to evaluate the performance of the algorithm. Our analysis shows that the proposed method is accurate and shows better clustering performance along with valuable insights for each cluster. Our approach is unsupervised, scalable and can be applied to various domains.**

Keywords: Self Organizing Map, K-Means, Genetic Algorithm, Clustering, Health care, Data Mining.

I. Introduction

Data mining and machine learning is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. The need for detecting unidentified valuable information in the medical data, forces the field of data mining to be more popular in the healthcare domain. Data mining offers many benefits in the health industry domain including, lower cost of available medical treatment for patients, the identity of different patients, exploring causes of various diseases, and determining the possible methods of treatment. It also helps health care researchers, to make efficient health care policies, constructing drug recommendation systems, and developing health profiles of different individuals [1].

Nowadays, health care data are received from various health care service providers to provide better health care services. The content of this data contains details of patient, medical tests, as well as their diagnosis and treatments. This data is usually very enormous and complex. So, it is considered to be challenging for analysis in order to take a decision in regards to a patients health. It can be difficult to synthesize data in a meaningful way especially when data is multilayered. This process needs specific yet powerful information technology tools. For data use in scenarios of healthcare monitoring systems for examples, cardio vascular patients, such ways of data mining and processing accurate and valid highlights requires the use of enormous groups of heterogeneous data [2].

Another application of data mining techniques is to provide benefits to healthcare organizations for segmenting and grouping patients that have similar type of diseases or health issues, so the organization can provide them with effective treatments [3]. In regards to patient health conditions, the decision by health care service providers could be determined by using several data mining techniques such as classification, clustering, and regression. One of the well-known techniques of data mining in health care data analysis is classification. Classification can predicate target with such precision within the data segment [4]. For example, a patient can be classified as “normal” or “infected” depending on the basis of patterns in the data. There are different classification methods such as k-nearest neighbor (KNN), decision trees (DT), support vector machine (SVM), and ensemble approaches used in classification techniques. The specific use of classification remains a learning method that is supervised within the class parameters [5].

Alternatively, clustering is classified as unsupervised learning technique which is used extensively in the field of data mining. Unlike classification, there is no predefined class categories in the method of clustering. For many years and over a span of experience, various clustering techniques are developed and used in many applications. Clustering is based upon dividing the large data set into smaller subsets or groups where each group has similar characteristic which can be measured [6]. What makes clustering effective for data handling purpose is the fact that it needs little or no identifying information about the data to complete the analysis. The algorithm K-Means clustering is considered one of the widely used data clustering methods in which data sets are partitioned into k groups or clusters. The grouping algorithm K-Means was originally composed by MacQueen [7] and later enhanced by Hartigan and Wong [8]. Although K-Means algorithm are

978-1-5090-0809-4/16/$31.00 copyright 2016 IEEE
SERA 2016, June 8-10, 2016, Baltimore, USA

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

extensively used in many places, there are several problem factors associated with it. The major problem is to specify the number of clusters in advance for K-Means clustering. Secondly, the initialization of cluster centroids sometimes leads to incorrect cluster output.

There are currently other existing frameworks in the literature that combine K-Means with Self Organizing Map (SOM) in the domains such as market segmentation, classification of sensor data, navigation patterns, etc. The concept of combining two methods is not novel, as there is various approaches combining two methods together. This includes the work of Haraty, et al where k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data [9]. Soliman et al., employed SVM classification in the classification of various diseases, SVM was also used together with k-means clustering and was applied to microarray data for the identification of the diseases [10]. Additionally, Tapia et al. analysed gene expression data aided by the hierarchical clustering approach by genetic algorithm [11]. However according to Tomar and Agarwals survey [5], there has not been a combination to create clusters using SOM with Genetic K-Means in two healthcare datasets.

This paper, proposes an algorithm which uses Self Organizing Map (SOM) method with Genetic K-Means clustering algorithm to mechanically discover the optimal number of sections in the data. SOM is known as a data visualization and clustering technique which is based on a neural network viewpoint [12]. The fundamental difference between SOM and K-Means is that SOM produces number of segments based on predetermined topographic ordering relationship. Throughout the training process, SOM uses every data point to renew the segments and produces an ordered set of centroids for any given data set in the end. In other words, SOM overcomes the limitation of K- Means algorithm by geometrically calculating the number of centroids in the data. SOM will specify the amount of k-cluseters, where centroids are based on GA, which makes sure the centroids don't overlap. The centroids from the different subspaces after SOM constructs different segments were calculated using R software where the GA is constructed within.

The rest of this paper is organized as follows. Section II presents a review of data mining techniques used in health care. Section III describes the proposed technique along with other clustering techniques used in this paper for comparison. Section IV provides a description of the dataset along with the experimental setup and Section V provides a discussion of the results as well as visualization of insights from clusters. Finally, Section VI concludes this paper along with future extension of the research.

II. Related Work

Health care involves detailed investigation of disease such as diagnosis, treatment and prevention, injury and other physical and mental impairments within human beings [13]. The health care industries in most countries are evolving at a rapid pace. It produces enormous amounts of data this includes electronic medical records, administrative reports, and other research findings [14]. However, those health care data are always being under utilized. As discussed earlier in this paper, data mining techniques are able to search for new and valuable information from these large volumes of data. Data mining in health care are mainly being used to predict various diseases as well as assisting the diagnosis in order for the doctors to make their clinical decision.

In order to identify a set of clusters or categories that describe the data, clustering is necessary as a common task for description [15]. Velosoa et al. [16] applies the method of vector quantization of K-Means to explore the readmissions in intensive medicine. In the statistical classifiers, Su et al. [17] implemented the Mahalanobis Taguchi System (MTS) for ulcers disease where the prediction model design is targeted. However, due to class imbalance problems, this method was unable to provide good results on the real unknown data set. Similarly, authors in [18] and [19] have used the linear discriminant analysis in their respective work. Jen et al [19] predicts the severity Parkinsons disease patient using scores of non-motor symptoms. This study focuses upon the quantifiable analysis of interactions between people with both motor and non-motor issues. Other studies by Laencina et al. [20], Zheng et al. [21] and Kang et al. [22] show a connection relation between the use of Support Vector Machines (SVM) where the data mining models are built to help in the diagnoses of the medical field.

The importance study of clustering is not limited to computer scientist; however, it is also valuable for patterns recognition experts and statisticians. Bradley et al. [23] develops an algorithm based on K-Means but due to sequential nature of algorithm, it was restricted by memory limitations. Aggarwal et. al [24] supports framework that looks to diffuse the tool of clustering, so that there are many steps within the process. There is a step within the statistical information storage that allows the next step to sum up how the collected information varies in traits per the k constant which represents the number of clusters. Their experimental shows promising results in both memory efficiency and accuracy. Arthur et al. [25] create a method to initialize K-Means via selecting a random centers at the beginning with a specific probabilities. Celebi et al. [26] investigate the development of K-Means algorithm while focusing in its initialization method. This study remains important as the details of the eight step process which defines initialization but this also calls into question other options as well. Himanshu et al. [27] implemented a clustering technique using singular vector decomposition to discover how many number of clusters is required. The authors applied the K-Means algorithm to create clusters. Furthermore, they refined these clusters by using feature voting while the phase of refinement allows the algorithm to outperform the conventional K-Means algorithm.

III. Methods

In this section, we describe traditional K-Means clustering algorithm, Genetic algorithm and our proposed SOM based Genetic K-Means algorithm in detail.

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

A. K-Means Clustering

K-Means is a clustering algorithm that comes under unsupervised learning methods. K-Means is used to find partitions inside the data [28] [29]. Given a set of data points $(x_1, x_2, \ldots, x_n)$, where each data point is a d-dimensional vector, K-Means clustering tried to segment the n observations into a set of k clusters $(\leq n)$ such as $S = S_1, S_2, \ldots, S_k$ with the focus on minimizing the within-cluster sum of squares (WSS) which can be described as sum of distance of each data point in the cluster to the k centers. The objective function of K-Means is to find

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x_i - c_i\|^2 \tag{1}$$

where c_i is the centroid of points in S_i.

B. Genetic Algorithm

Use of genetic algorithms fall under a larger standard of algorithms called evolutionary algorithms or EA. These EAs work to promote problem solving toward optimizing techniques which promote significant variance in genetic features like inheritance, mutation, selection and crossover of attributes [30]. Much of what takes place as random with the population of user/individuals and this can be labeled as a generational group. To evaluate, each group generation is assessed for fitness within the optimization for serving to promote problem solving. The more up to the challenge the individual proves to be within the population, the more signifies the rate of individual genome mutation which leads to creating the next generation of genetic characteristics. Thus, the algorithm continues to provide solutions within the generational testing. With the maximum amount of generations produced during the process, the algorithm finalizes and the fitness level standard is reached.

C. DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is designed to discover arbitrary shaped segments in the data [31]. In DBSCAN, the density associated with a point is obtained by counting the number of points in a region of specified radius around the point. Here clusters can be defined as a specific threshold for points within density. DBSCAN also means a value for radius or as the user defines the distance measure or a value MinPts. This can be minimal amount of points seen in Eps ϵ radius. Choosing the DBSCAN algorithm allows our study to be profound comparison when looking at patterns within clusters. It is useful as an input algorithm.

D. Self Organizing Map (SOM)

There is a concept called the Self Organizing Map or SOM which is also referred to as the Kohonen map. This is another type of algorithm that is derived from unstructured artificial neural network. This map is a concrete and valuable tool for organizing clusters within complex data [12]. The SOM allows for a nonlinear design of mapping with the high dimensional context of data space. This mapping accounts for topological planning and preservation. A map contains m unites (or neurons) which is set on a regular grid of low dimensions to defines the relationship of its neighborhood.

In mathematical context, let $x = (x_1, x_2 \ldots x_p)$ be the input vector and $w_l = (w_{l1} w_{l2} \ldots w_{lp})$ be the weight vector associated with the node l where w_{lj} indicates the weight assigned to input x_j to the node l and p is the number of input variables. In random order, each of the data set objects is introduced to the network. Kohonen's learning law is an algorithm which determines the closest node to each of training case and shifts the winning node to be much closer into the training case. The distance between the node and the training case is moved some proportion which is managed by learning rate. For each object i in the training data set, the distance d_i between the weight vector and the input signal is calculated. Then the competition starts and the node with the smallest d_i is the winner. The weights of the winner node are then updated using some learning rule. The weights of the non winner nodes are not changed. Although any metric could be selected, the comparison of each node with each object is made by Euclidean distance. The Euclidean distance between an object with observed vector $x = (x_1, x_2 \ldots x_p)$ and weight vector $w_l = (w_{l1} w_{l2} \ldots w_{lp})$ is given by

$$d(x, w_l) = \left(\sum_{j=1}^{p} (x_j - w_{lj})^2 \right)^{\frac{1}{2}} \tag{2}$$

Let w_s^l be the weight vector for the l^{th} node on the s^{th} step of the algorithm, x_i be the input vector for the i^{th} training case, and α^s be the learning rate for the s^{th} step. On each step, a training case x_i is selected and the index q of the winning node (cluster) is determined by

$$q = \arg\min_{l} \|w_s^l - x_i\| \tag{3}$$

The weight vector closest to the input vector is called Best Matching Unit (BMU). The Kohonen weight update rule for the winner node is given by

$$w_q^{s+1} = w_q^s(1 - \alpha^s) + x_i = w_q^s + \alpha^s(x_i - w_q^s) \tag{4}$$

After the training phase, SOM constructs different topological segments based on BMU where each BMU represents one segment.

E. Proposed Algorithm

We propose an efficient K-Means clustering algorithm which uses the SOM method to discover the optimal segments number in the data as a preprocessing step. The main feature of SOM revolves on transforming the input data into topological features to be preserved on the map. Once the data is transformed and optimal number of segments are calculated, the transformed data is passed into Genetic K-Means algorithm to calculate the final clustering results. Figure 1 shows the flow of our algorithm.

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

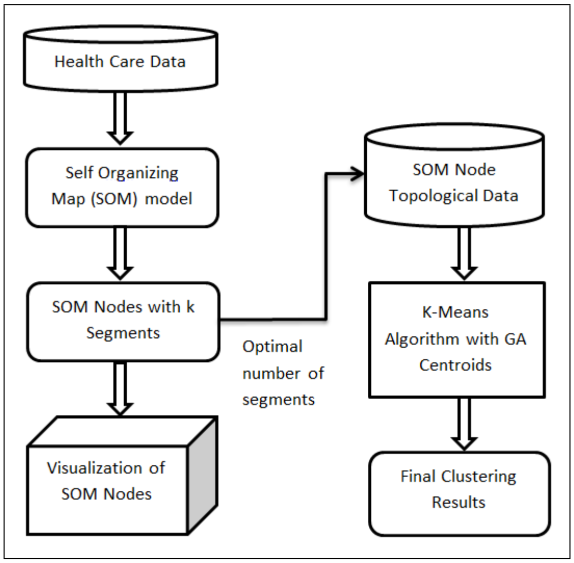
Health Care Data
Self Organizing Map (SOM) model
SOM Nodes with k Segments
Visualization of SOM Nodes
Optimal number of segments
SOM Node Topological Data
K-Means Algorithm with GA Centroids
Final Clustering Results

Fig. 1: Efficient SOM Genetic K-Means Clustering Flow Diagram.

The figure shows the process of efficient SOM Genetic K-Means clustering algorithm (a) SOM algorithm to get optimal number of segments (b) Genetic algorithm to get optimal centroids (c) K-Means algorithm to get clusters (d) Visualization stage to get the valuable insights from data.

The algorithm starts by taking an input data set and processing it using SOM model. Inside the SOM model, the algorithm initializes all node weight vectors randomly and finds the “Best Matching Unit” (BMU) in the map to get the most similar node together. Similarity is calculated using the euclidean distance formula for all the nodes within the neighborhood of the BMU. Since SOM is an iterative algorithm, it adjusts weights of nodes in the BMU with respect to learning rate until the solution converges and weight vectors becomes constant.

Once the conversion is made, the number of BMU indicates the optimal number of segments in the data. Our proposed algorithm goes to the next stage and uses the number of segments to specify the input parameter k as the number of clusters to Genetic K-Means clustering algorithm. Once Genetic algorithm outputs the greatest centroids, K-Means algorithm is based on those centroids and the clustering results obtained. At this point, visualization is used to develop useful insights from each cluster and report the analysis. The pseudocode of the algorithm is stated in Algorithm 1.

IV. Experimental Evaluation

A. Dataset Description

This section, describes the datasets and experimental setup used in this paper. We used two datasets on liver disease and heart disease from UCI Machine Learning Repository [32]. Liver disease dataset comes from BUPA Medical Research Ltd which contains seven variables of blood tests which are considered to be sensitive to liver disease that could arise from excessive alcohol consumption. The liver dataset is comprised of 345 patients who also contain a class label if a patient has a liver disease or not. The details for this dataset can be found in Table I. Heart disease dataset is collected from Cleveland Clinic Foundation and contains fourteen variables concerning heart disease diagnosis. The heart dataset contains data from 303 different patients in addition to the class label if the patient is going through any heart disease. Table II explains each variable of heart disease data in detail.

Algorithm 1 Efficient SOM Genetic K-Means Algorithm

Input: Input dataset D with n features, size of grid W with i and j as dimensions, learning rate α

Output: Output dataset with k cluster labels

```
procedure SOM –Genetic K–Means–
    while $\alpha \geq 0$ do
        for each $x \in D$ do
            for each $w_{ij} \in W$ do
                Calculate $d_{ij} = \|x - w_{ij}\|$
                Select $BMU$ that minimizes $d_{ij}$
                Update each weight vector $w_{ij} \in W$
                Decrease $\alpha$
            end for
        end for
    end while
    Intermediate Outputs: (i) SOM Topological Data $TData$ (ii) Optimal number of clusters $k$
    centroids = GA-Centers(TData, $k$)
    clusters = K-Means(TData, centroids)
end procedure
```

TABLE I: Dataset Description of Liver Data set

Liver Disease Dataset	
Attribute Name	Description
mcv	mean corpuscular volume
alkphos	alkaline phosphotase
sgpt	alamine aminotransferase
sgot	aspartate aminotransferase
gammagt	gamma-glutamyl transpeptidase
drinks	alcoholic beverages drunk per day
selector	class label for liver disease

Attribute shows the feature names and Description explains each feature of Liver disease data set.

B. Performance Metrics

Both datasets include a variable that contributes to the final diagnosis of each individual patient. The performance of the clustering algorithm was evaluated by computing metrics related to classification accuracy and visual insights to show the correctness of the results. Using class variables for each patient, we determine a weighted average accuracy. The assignment of each cluster to the class is based on the most frequent variable within the cluster. Then, the accuracy of this assignment is evaluated by calculating the percent of correctly assigned records [33]. The accuracy reported is averaged through all clusters and are weighted by the number of records in each cluster. We also create some visualization that shows

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

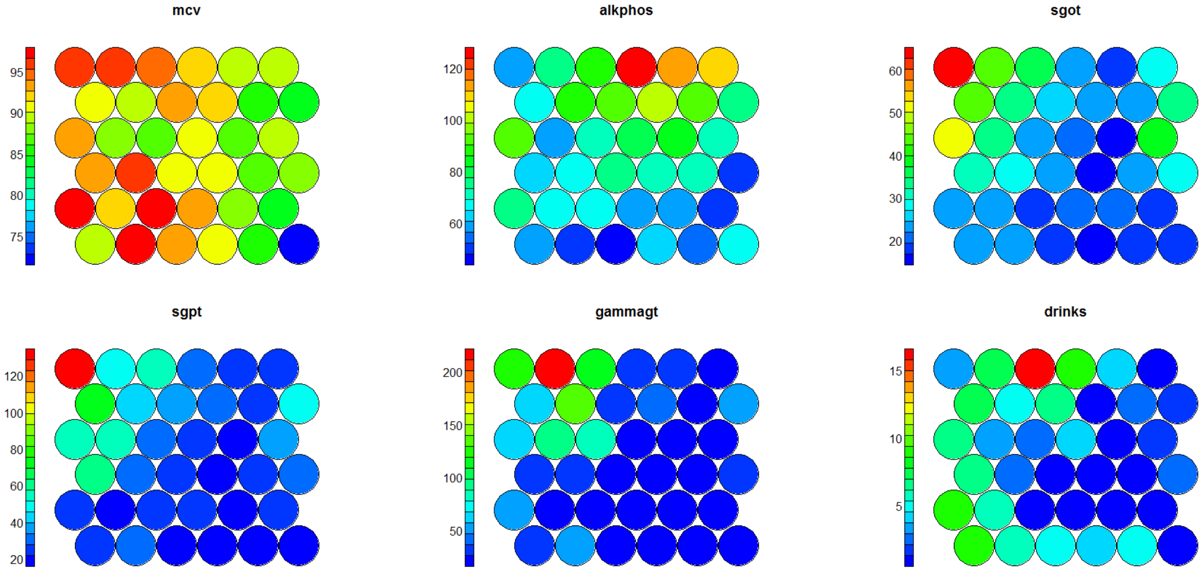

Fig. 2: Heatmaps for visualisation of possible Self-Organizing Maps

The figure shows the visualization of SOM for each variable in Liver disease dataset. Each plot in is scaled to their real values in order to see the correct relationships between variables. The color bar on y-axis shows the min and max value for each feature.

TABLE II: Dataset Description of Heart Data set

Heart Disease Dataset	
Attribute Name	Description
age	age in years
sex	patient gender
cp	chest pain type
trestbps	resting blood pressure
chol	serum cholestoral
fbs	fasting blood sugar
restecg	resting electrocardiographic results
thalach	maximum heart rate
exang	exercise induced angina
oldpeak	ST depression
slope	he slope of the peak exercise ST segment
ca	number of major vessels
thal	exercise test
num	diagnosis of heart disease

Attribute shows the feature names and Description explains each feature of Heart disease data set.

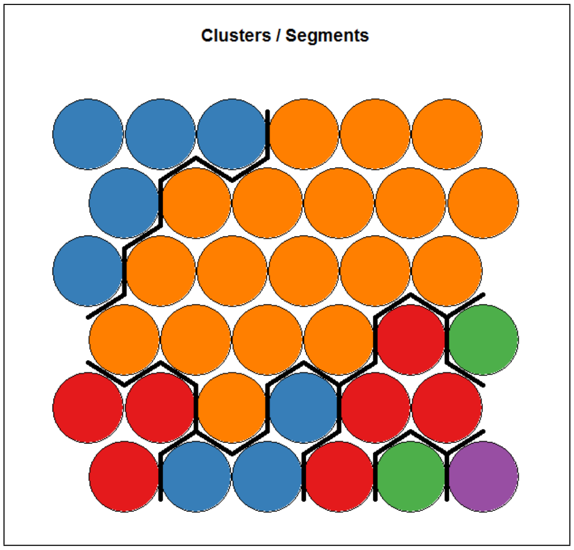

Fig. 3: Cluster representation of liver disease data using SOM Genetic K-Means algorithm.

The figure shows the visualization of different clusters generated by SOM Genetic K-Means clustering algorithm using liver disease data when number of clusters are set to 5. Each color represents a different cluster/segment.

the key findings in each cluster and help professional to make use of it.

V. Results and Discussions

A. Cluster Representation

A set of experiments were preformed during each stage of our proposed algorithm. Figure 2 shows the SOM segments generated in the first stage using liver disease data. The feature size was set (number of neurons) as 36 which can be seen as 6 x 6 grid in the heatmap. These 36 number of neuron are selected based on the statistics provided by other researchers where they estimate that each neuron can be assigned to 10 data points [33]. In liver disease dataset, we have 345 data points, so dividing total data points with 10 gives us 35 as a rounding number and the closest squared number is 36. Therefore, we picked 36 number of neurons as our input. SOM algorithm takes each variable separately and finds the optimal number of segments/clusters. We identify 5 clusters and are represented by different colors when looking at the heatmaps for all variables in Figure 2. An insight to be taken

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

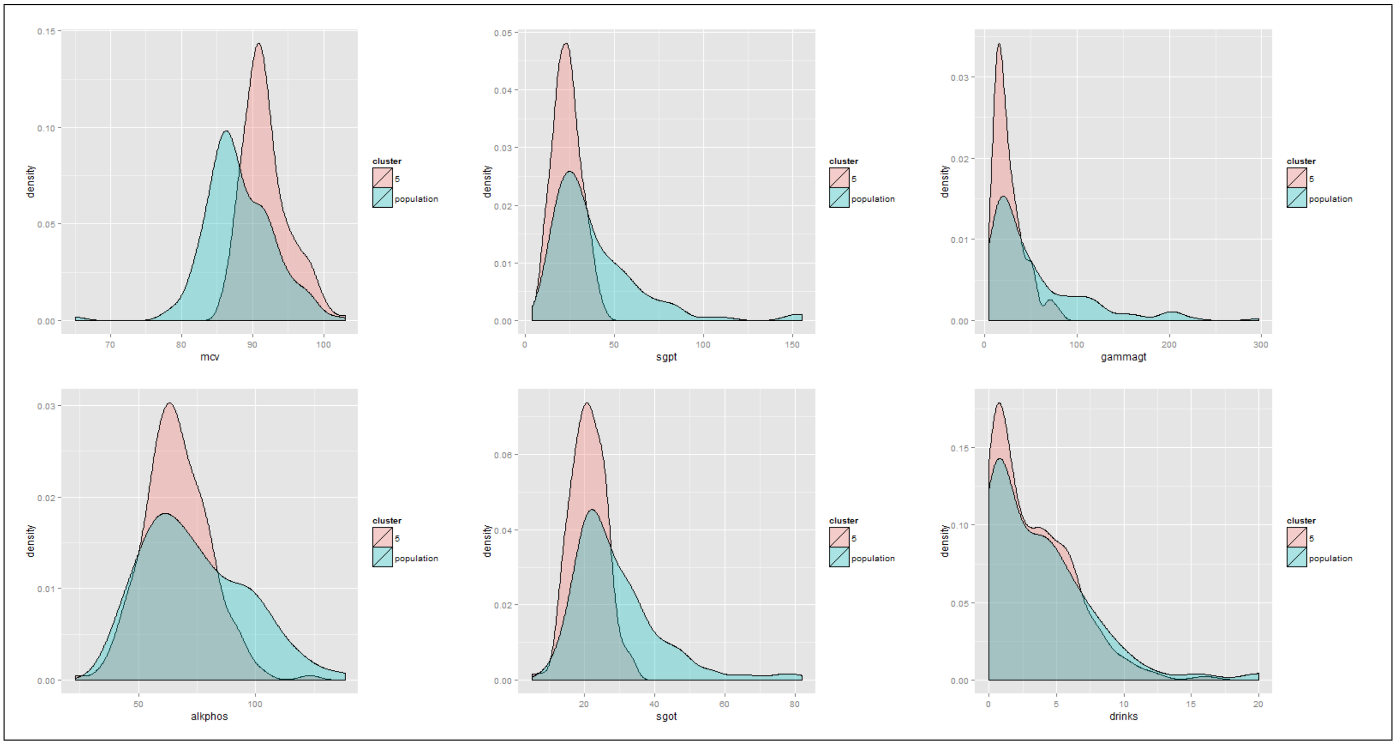

Fig. 4: Density plot using liver disease data to get insights on cluster 5

The figure shows the comparative density plot of a cluster 5 for each feature separately against other clusters i.e whole population using liver disease data.

from the heatmap is the relationship between each variables. For example, it is noteworthy that alkphos and sgpt variable heatmaps show an inverse relationship in many places of heatmap; therefore, we can infer that patients having low value of sgpt usually have high value of alkpos and vice versa. Further all heatmaps are visualized side by side and can be used to build up a picture of the different variables and their characteristics.

After the first stage, Genetic K-Means algorithm was run on SOM Topological data with optimal number of clusters k found in first stage. Figure 3 shows the final clustering output when the numbers of clusters are set to 5. As can be seen from the figure that there is five different clusters represented by different colors and separated using cluster boundaries. The same process was repeated for heart disease dataset and 4 clusters were identified by SOM algorithm.

B. Cluster Evaluation

As explained in Section IV-B, a weighted average accuracy was determined for the clustering output using class labels. For example, in the liver disease dataset, the variable "selector" is the class which tells whether patient is normal or infected. To evaluate each cluster, we presume that each cluster represents the most frequent class present in it, and then the accuracy of this assignment is evaluated by computing the percent of correctly assigned records. The reported accuracy is averaged across all clusters, weighted by the number of records in each cluster [33]. The mathematical representation of weighted average accuracy is as follows

TABLE III: Performance Evaluation using Classification Accuracy

	Weighted Classification Accuracy (%)		
Dataset	SOM Genetic K-Means	K-Means	DBSCAN
Liver Disease	**73.84**	69.15	67.66
Heart Disease	**69.90**	66.27	61.45

The number shows the weighted classification accuracy of different algorithms in percentage using liver and heart disease datasets.

$$\sum_{i=1}^{k}(n_i * Accuracy))/N \tag{5}$$

where n_i is the number of data points in cluster i, $Accuracy$ is the classification accuracy for cluster i and N is the total number of data points in the dataset.

Table III shows the classification accuracy for our proposed method and various other methods based on two different datasets. Our empirical shows that the proposed algorithm outperforms other methods in terms of classification accuracy and is able to generate useful insights. When comparing the proposed method with other existing clustering methods that work on healthcare datasets, we are able to evaluate the proposed method. The proposed method combines two clustering methods together which makes the performance and computational efficiency very high in comparison to other clustering method with healthcare dataset.

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

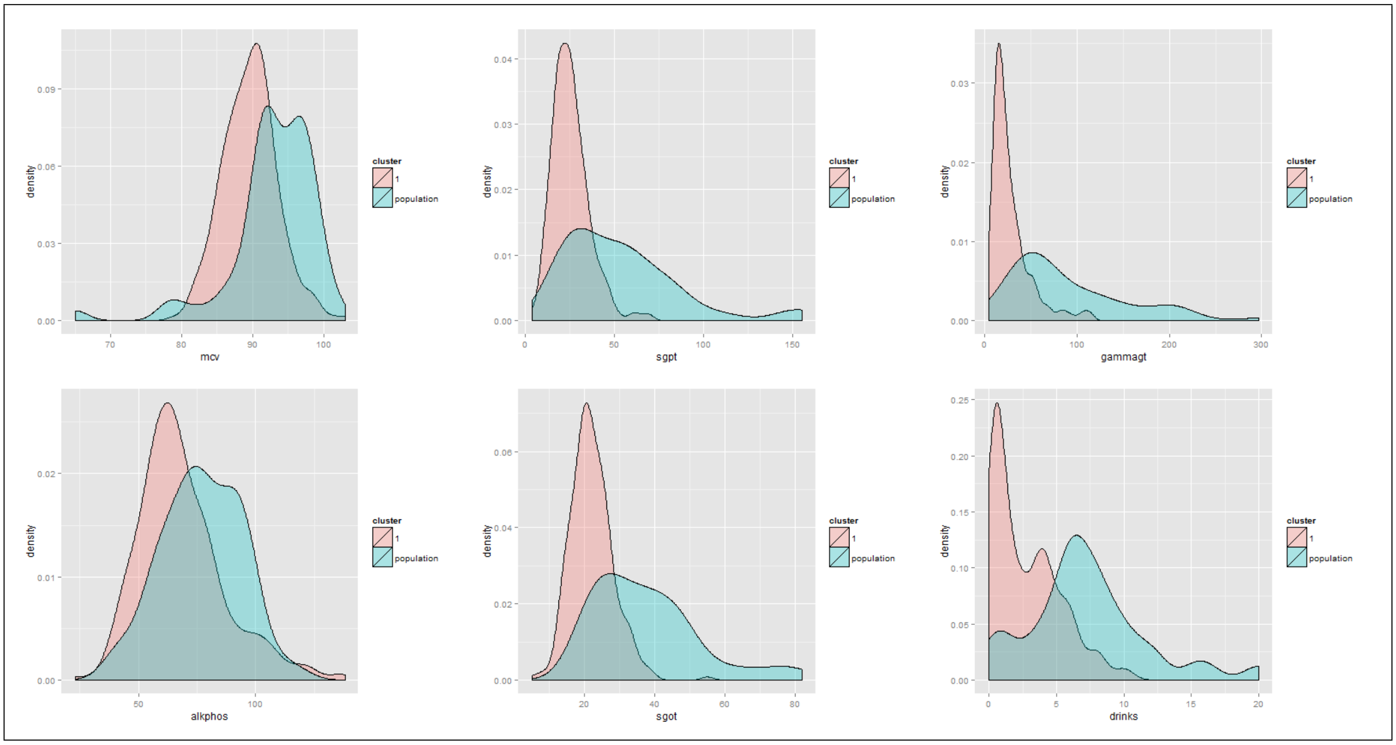

Fig. 5: Density plot using liver disease data to get insights on cluster 1

The figure shows the comparative density plot of a cluster 1 for each feature separately against other clusters i.e whole population using liver disease data.

C. Visual Insights Application

In order to provide a visual insight application in our proposed method, we processed our clustering results to gain more insights. Figures 4 and 5 show a comparative density plot for cluster 5 and cluster 1 in liver disease dataset. The idea behind this plot is to show how each cluster preserves its uniqueness as compared to other clusters. For example, Figure 4 shows that patients in cluster 5 have high mean value of “mcv” with respect to other clusters (whole population). Similarly, the uniqueness of cluster 5 in comparison to other variables is clear. Thus, we can say that these visualizations help professional to create interesting stories that best explain the variables in the dataset. Comparison between Figures 4 and 5 is done to see how cluster 1 patients shows lower mean on the variable of “mcv” while cluster 5 patients shows higher mean with respect to the whole population for the same variable. Thus, cluster 1 contains patients, which mostly have low ”mcv” value with respect to mean value of overall population.

VI. Conclusion

The paper proposed an effective method used to cluster and explore health care data. The method uses Self Organizing Map (SOM) method to infer optimal number of clusters at the initial stage, then apply Genetic K-Means algorithm to cluster the health care data for knowledge discovery. The empirical evaluation shows that our proposed algorithm outperforms other existing methods. It also presents a visual application using density distributions of the training variables in every cluster in order to build a meaningful picture of the cluster characteristics. For future work, we like to extend the algorithm and use hybridization to improve the performance. Additionally the method is to be used to solve other data mining challenges such as document clustering, image clustering etc. The algorithm is to be compared with other state-of-the-art methods.

Acknowledgment

The authors would like to thank the Department of Computer Science at Bowie State University.

References

[1] Hian Chye Koh, Gerald Tan, et al. Data mining applications in healthcare. *Journal of healthcare information management*, 19(2):65, 2011.

[2] Mary K Obenshain. Application of data mining techniques to healthcare data. *Infection Control & Hospital Epidemiology*, 25(08):690–695, 2004.

[3] Neesha Jothi, Wahidah Husain, et al. Data mining in healthcare–a review. *Procedia Computer Science*, 72:306–313, 2015.

[4] Cristoph Helma, Tobias Cramer, Stefan Kramer, and Luc De Raedt. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. In *J. Chem. Inf. Comput. Sci*, volume 44, pages 1402–1411, 2004.

[5] Divya Tomar and Sonali Agarwal. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013.

[6] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*, 2011.

[7] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.

[8] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[9] Ramzi A. Haraty, Mohamad Dimishkieh, and Mehedi Masud. An enhancedk-means clustering algorithm for pattern discovery in healthcare data. *International Journal of Distributed Sensor Networks*, 2015:1–11, 2015.

[10] Taysir Hassan A. Soliman, Adel A. Sewissy, and Hisham AbdelLatif. A gene selection approach for classifying diseases based on microarray datasets. *2nd International Conference on Computer Technology and Development(lCCTD 2010)*, 2010.

[11] J. J. Tapia, Enrique Morett, and Edgar E. Vallejo. A clustering genetic algorithm for genomic data mining. *Foundations of Computational Intelligence*, 4:249–275, 2009.

[12] T. Kohonen. *Self-Organizing Maps*, volume 30. Springer, ISBN 3540679219, 2001.

[13] Ji-Jiang Yang, Jianqiang Li, Jacob Mulder, Yongcai Wang, Shi Chen, Hong Wu, Qing Wang, and Hui Pan. Emerging information technologies for enhanced healthcare. *Computers in Industry*, 69:3–11, 2015.

[14] Christo El Morr and Julien Subercaze. Knowledge management in healthcare. *Handbook of research on developments in e-health and telemedicine: Technological and social perspectives*, pages 490–510, 2010.

[15] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[16] Rui Veloso, Filipe Portela, Manuel Filipe Santos, Álvaro Silva, Fernando Rua, António Abelha, and José Machado. A clustering approach for predicting readmissions in intensive medicine. *Procedia Technology*, 16:1307–1316, 2014.

[17] Chao-Ton Su, Pa-Chun Wang, Yan-Cheng Chen, and Li-Fei Chen. Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients. *Journal of medical systems*, 36(4):2387–2399, 2012.

[18] Rubén Armañanzas, Concha Bielza, Kallol Ray Chaudhuri, Pablo Martinez-Martin, and Pedro Larrañaga. Unveiling relevant non-motor parkinson's disease severity symptoms using a machine learning approach. *Artificial intelligence in medicine*, 58(3):195–202, 2013.

[19] Chih-Hung Jen, Chien-Chih Wang, Bernard C Jiang, Yan-Hua Chu, and Ming-Shu Chen. Application of classification techniques on development an early-warning system for chronic illnesses. *Expert Systems with Applications*, 39(10):8852–8858, 2012.

[20] Pedro J García-Laencina, Pedro Henriques Abreu, Miguel Henriques Abreu, and Noémia Afonoso. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in biology and medicine*, 59:125–133, 2015.

[21] Bichen Zheng, Sang Won Yoon, and Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482, 2014.

[22] Seokho Kang, Pilsung Kang, Taehoon Ko, Sungzoon Cho, Su-jin Rhee, and Kyung-Sang Yu. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. *Expert Systems with Applications*, 42(9):4265–4273, 2015.

[23] Paul S Bradley, Usama M Fayyad, Cory Reina, et al. Scaling clustering algorithms to large databases. In *KDD*, pages 9–15, 1998.

[24] Charu C Aggarwal, Jiawei Han, Jianyong Wang, and Philip S Yu. A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pages 81–92. VLDB Endowment, 2003.

[25] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[26] M Emre Celebi, Hassan A Kingravi, and Patricio A Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

[27] Himanshu Gupta and Rajeev Srivastava. k-means based document clustering with automatic k selection and cluster refinement. *International Journal of Computer Science and Mobile Applications*, 2(5):7–13, 2014.

[28] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[29] John Burkardt. K-means clustering. *Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics*, 2009.

[30] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Genetic algorithm-based clustering technique. *Pattern recognition*, 33(9):1455–1465, 2000.

[31] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[32] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[33] Erzsébet Merényi, Michael J Mendenhall, and Patrick ODriscoll. Advances in self-organizing maps and learning vector quantization.

Authorized licensed use limited to: Jerusalem College of Technology. Downloaded on February 01,2025 at 17:17:37 UTC from IEEE Xplore. Restrictions apply.