International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Improve of Fuzzy C-Means Clustering in Feature Extraction Phase on the Breast Cancer Analysis

A. Josekin[1], D. Sudhakar[2]

[1]Scholar, Department of Computer Science ,CSI Bishop Appasamy College of Arts and Science, Race Course, Coimbatore, Tamil Nadu, India

[1]Assistant Professor, Department of Computer Science, CSI Bishop Appasamy College of Arts and Science, Race Course, Coimbatore, Tamil Nadu, India

ABSTRACT

Cancer analysis is one of the broadly advised acreage in the healthcare domain. The objective of the breast cancer problem is to predict the property of a new tumor (malignant or benign). The existing method hybridizes K-means algorithm and SVM (K-SVM) for breast cancer diagnosis. To reduce the high dimensionality of feature space, it extracts abstract malignant and benign tumor patterns separately before the original data is trained to obtain the classifier. In order to improve the quality of prediction, Fuzzy c-means clustering is hybridizes with SVM (F-SVM). An improved fuzzy c-means algorithm is proposed to deal with the cancer data. The proposed algorithm improves the traditional Fuzzy c-means algorithm in terms of selecting the initial cluster centre. Thereby, it avoids the basic drawback of Fuzzy C-means and improves the quality of prediction over k-means algorithm. It helps to predict the benign and malignant tumors. Based on the derived membership, each tumor pattern is considered as a model. Further support vector machine (SVM) technique is used to obtain the new classifier to discriminate the data.

Keywords : K-means, Fuzzy C-means, K-SVM, F-SVM.

I. INTRODUCTION

The research ideas in Data mining technology has seen an increasing popularity in a variety of fields while Medical Data mining has emerged to be the current area of active research. Mining in this field holds huge significance providing deeper knowledge and diagnosis of diseases like diabetes, stroke, disorders, cancer and heart disease. In fact, heart disease, cancer and liver disorders are highlighted in this research work, as these have become a leading reason behind mortality in all age groups in the recent years Consequently, the World Health Organization (WHO) mentioned that 30% of deaths are because of heart disease or Cardio Vascular Disease (CVD), and 45% of deaths are due to cancer implications and 10% due to liver disorders. Many numbers of mining tools are available for the accurate diagnosis with medical data, and among these clustering is considered to be reliable for predicting with accuracy in this task. But, along with the benefits obtained from this method, there are also few challenges such as lack of precision and quality. The lack of precision in a dataset is frequently activated by restricted observation, constrained perception and less amount of resources in the collection of data. In order to prevail over the results associated with the process of mining, the optimal feature selection methodologies with modified distance measure are investigated in this study.

An interesting trend of research work is to accomplish the data mining strategies in order to do the feature selection before deriving predictive models according to the statistical approaches like logistic and Cox regression. For instance, these approaches were enforced to derive prognostic survival models after thoracic transplantations and in the ICU. Furthermore, to highlight the crucial variable subspaces while supervising the blood glucose of ICU patients, subgroup discovery methods should be utilized. The most appealing feature of the research on biomarker discovery, like the work of Baumgartner et al.,(who examined a huge set of mass spectrometry data to drew-out the metabolites which is appropriate for infarction) deal with the problem of feature selection. In order to choose the variable which is helpful for intermediate staging of renal cancer that begins from tissue microarray data, so Cox regression and a bootstrap elimination scheme were utilized.

II. LITERATURE REVIEW

The invention in data mining has becoming popular in numerous areas in which Medical Data mining is the dynamic research area nowadays, which gives in-depth knowledge and analysis of diseases for instance diabetes, disorders, stroke, cancer and heart disease. This chapter evaluates the methods exist in the research literature for feature selection, importance of clustering algorithms and gene expression dataset analysis of cancer.

2.1 Feature Selection Algorithms

The main problem in medical diagnosis of data is feature selection. Features are selected depends on the property of the data. It deals with the redundant and irrelevant data**.**

A multiple Support Vector Machine Recursive Feature Elimination (SVM-RFE) for gene selection in cancer classification along with expression data was presented Duan et al., (2011). It recommended an innovative feature selection technique, which uses a backward elimination technique close to that used in SVM-RFE. The introduced technique was dissimilar from the SVM- RFE technique, since it assessed the feature ranking score from a numerical investigation of weight vectors of various linear SVM, which is trained on subsamples of the actual training data in all the steps.

Feature selection is mainly used for selecting the best attributes from the given data especially in medical diagnosis. Heuristic methods help to resolve the problem of selecting best features described by Subanya and Rajalaxmi (2014). ABC is a metaheuristic algorithm that share information between the bees in the population and select feasible solutions, which can satisfy the defined criteria. ABC has a unique solution update mechanism (updating in two phases) which allows the results to converge to the optimal solution quickly. Swarm intelligence based Artificial Bee Colony (ABC) algorithm has been proposed to find the best features in the disease identification. To evaluate the fitness of ABC, SVM classification is used. The performance of the proposed algorithm is validated against the Cleveland Heart disease dataset taken from the UCI machine learning repository with the 303 samples. The experimental results showed that ABC–SVM performs better than forward feature selection with reverse ranking. The results also showed that the designed method obtained good classification accuracy with only seven features.

Walaa Gad (2016) designed a method to improve the diagnosis of breast cancer on WDBC and Wisconsin Prognosis Breast Cancer Dataset (WPBC), in which he combined an unsupervised learning method K-means with SVM a supervised learning method. This method eliminates the inapplicable attributes using feature selection method chi-square. Method also improves the performance by speeding up and also eliminates the curse dimensionality.

2.2 Importance of clustering Algorithms

Fuzzy k-c-means clustering technique is utilized for medical image segmentation that was presented in Ajala, 2012. Fuzzy-c-means is a clustering algorithm that lets one part of data be in two or more clusters and k-means is an easy clustering technique wherein we utilize low computational complexity as contrasted to fuzzy c-means. Both Clustering techniques were joined to create a time effective segmentation technique known as fuzzy-k-c-means clustering method. They presented thresholding that is the most basic method for medical image segmentation, wherein this technique splits pixels in diverse classes based on their gray level. It moves toward division of scalar images by creating a binary partition of the intensity values of an image and identifies an intensity value. This intensity value is named as threshold that splits the desired classes. Classifiers are called supervised techniques as they need training data that are physically segmented and after that utilized it for mechanically segmenting novel data.

III. METHODOLOGY

3.1 Fuzzy C Means clustering

The Fuzzy c-means clustering algorithm is one of the earliest clustering approaches. The first version of this algorithm performed a hard cluster partition. This way, a data point could be a member of only one cluster. But in order to treat data belonging to several clusters, a fuzzy version of this algorithm was introduced by (Dunn, 1973). Later it was generalized by (Bezdek, 1973) thus producing the final version with the introduction of fuzzifier m. This final c-means algorithm recognizes spherical clouds of points in a p-dimensional space. Each cluster is assumed to have similar sizes and represented by its centre. Euclidean distance between a data point and center is used. This algorithm uses predetermined number of clusters, but it does not make an optimization on number of clusters.

$$0 = \frac{\partial}{\partial k_i} J = \frac{\partial}{\partial k_i} \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^m \|x_j - k_i\|^2$$

$$= \sum_{j=1}^{n} (\mu_{ij})^m \frac{\partial}{\partial k_i} \|x_j - k_i\|^2$$

$$= \sum_{j=1}^{n} (\mu_{ij})^m \lim_{t \to 0} \frac{\|x_j - (k_i + t\xi)\|^2 - \|x_j - k_i\|^2}{t}$$

$$= \sum_{j=1}^{n} (\mu_{ij})^m \lim_{t \to 0} \frac{1}{t} \Big(\big((x_j - k_i) - t\xi\big)^T \big((x_j - k_i) - t\xi\big) - (x_j - k_i)^T (x_j - k_i) \Big)$$

$$= \sum_{j=1}^{n} (\mu_{ij})^m \lim_{t \to 0} \frac{-2t(x_j - k_i)^T \xi + t^2 \xi^T \xi}{t}$$

$$= -2 \sum_{j=1}^{n} (\mu_{ij})^m (x_j - k_i)^T \xi$$

then since,

$$\frac{\partial}{\partial k_i} J = 0$$

$$\Leftrightarrow \sum_{j=1}^{n} (\mu_{ij})^m (x_j - k_i) = 0$$

$$\Leftrightarrow k_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m x_j}{\sum_{j=1}^{n} (\mu_{ij})^m}$$

This Fuzzy c-means algorithm works by assigning membership to each point corresponding to each cluster center using the distance between the cluster center and the data point. This membership takes a value between 0 and 1 and it shows the probability that a particular data point falls into a certain cluster. If the data point is closer to the center of the cluster, the membership value gets a high value. Since, the membership shows the probability that a particular data point falls into a certain cluster, the summation of the membership values for each data is equal to unity. This algorithm is composed of iterations and after each iteration, membership and cluster centers are updated according to the following formula.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{ik}})^{(2/m-1)}}$$

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m}, \forall j = 1,2,\ldots c$$

Where, n is the number of data points, vj represents the jth cluster center, m is the fuzziness index m ∊ [1,∝], c represents the number of cluster center. μ_{ij} represents the membership of ith data to jth cluster center, d_{ij} represents the Euclidean distance between ith data and jth cluster center.

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \|x_i - v_j\|^2$$

Where, ‖ xi – vj ‖ is the Euclidean distance between ith data and jth cluster center.

3.2 F-SVM Algorithm

The Fuzzy c-means algorithm is used for clustering tumors based on similar malignant and benign tumor features respectively. Further, it can be used as an optimization technique to minimize the overall distance between cluster centroids and cluster members. It is important to normalize the data point for eliminating the effect of the different feature scales. To train the centroids used to construct the cluster, the Fuzzy c-means algorithm repeatedly adapts the centroid location for reducing the Euclidean distance.

The cluster center is considered as a new symbolic tumor of that cluster. In this case, the scale of the original data set has been reduced by the symbolic objects labeled by malignant and benign. The different trials for benign and malignant tumors are shown in Fig. 1. From the curve, three is chosen for the number of clusters for benign and malignant tumor sets separately since it is the local minimum in the range of K from two to 30.Each cluster represents a specific tumor pattern. Each cluster centroid symbolizes the symbolic tumor of that cluster. To show the patterns clearly, the patterns for benign and malignant tumors are projected on three dimensions. The different color codes index different clusters.

After recognizing the malignant and benign tumor patterns, several symbolic tumors have been formed in both the malignant and benign data sets. The similarity between the untested tumor and the symbolic tumors plays an important role for diagnoses. The new feature is different from the previous one that contains only one feature; it is a profile tumor pattern, which are condensed information combining different previous features. Thus, the feature space dimension is reduced. The value of the new feature represents the similarity between the tumor and the pattern. The boundary between benign and malignant tumors is determined by training the SVM based on these new features.

3.3 F-SVM Clustering Algorithm

Let X = {$x_1, x_2, \ldots, x_n$} be the set of data points and V = {$v_1, v_2, \ldots, v_c$} be the set of cluster centers

Step 1: c cluster centers are selected randomly

Step 2: Calculate values (probabilities that a particular data point falls into a certain cluster) using the following formula

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{(2/m-1)}}$$

Step 3: Compute the Fuzzy c-means using

$$v_j = \frac{\sum_{i=1}^{n}(\mu_{ij})^m x_i}{\sum_{i=1}^{n}(\mu_{ij})^m}, \forall j = 1,2,\ldots c$$

Step 4: Repeat step 2 and 3 until the minimum J(U,V) value is reached.

IV. Performance Evaluation

To implement the method for this research, a data set of Wisconsin Diagnostic Breast Cancer (WDBC) from the University of California – Irvine repository has been used. The objective of the breast cancer problem is to predict the property of a new tumor (malignant or benign). The existing method hybridizes K-means algorithm and SVM (K-SVM) for breast cancer diagnosis. To reduce the high dimensionality of feature space, it extracts abstract malignant and benign tumor patterns separately before the original data is trained to obtain the classifier. In order to improve the quality of prediction, Fuzzy c-means clustering is hybridizes with SVM (F-SVM). Therefore, this section describes the results and comparisons between K-SVM and F-SVM. The following figure 1 shows the classification of breast cancer data as malignant and benign based on the feature selection method.

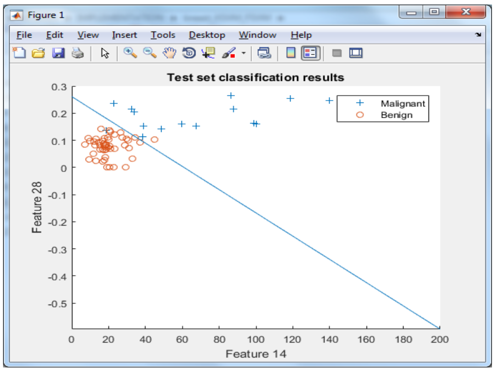

Figure 1: Classification of Breast Cancer Data

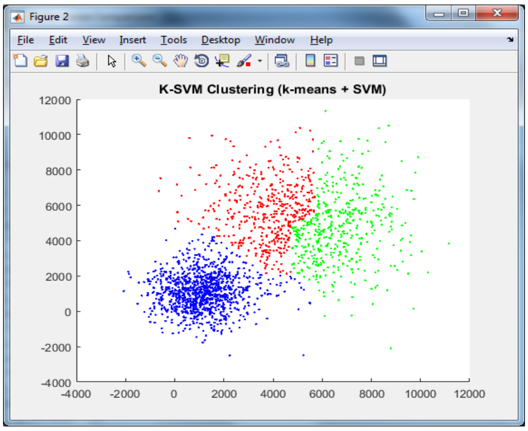

Figure 2: Clustering of Data using K-SVM Clustering

The Figure 2 shows clustering of data using K-SVM method. The figure further illustrates that three clusters are formed in the given breast cancer dataset. Similarly, the following Figure 3 illustrates the data clustering using F-SVM method, which is a newly proposed algorithm to cluster the given breast cancer dataset.

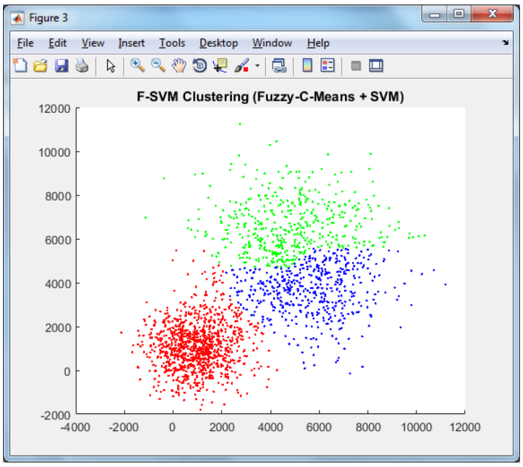

Figure 3: Clustering of Data using F-SVM Method

F-SVM technique is proposed in this thesis, which is a hybridization of Fuzzy c-means and SVM technique. The proposed method is compared with an existing method known as K-SVM. The primary evaluation parameters are time consumption and cluster efficiency. The following Figure 4 depicts the time consumption pattern of both algorithms, which states that F-SVM algorithm outperformed than K-SVM algorithm. Similarly, the cluster efficiency is estimated from the confusion matrix, which is shown in the Figure 5. The result states that K-SVM method has an efficiency of 90%, whereas the proposed F-SVM method has an efficiency of 96%. Therefore, the proposed method helps to achieve more efficiency than the existing method.

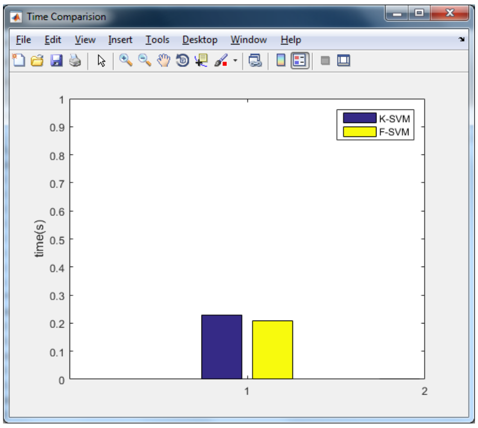

Figure 4: Time Consumption of K-SVM and F-SVM Techniques

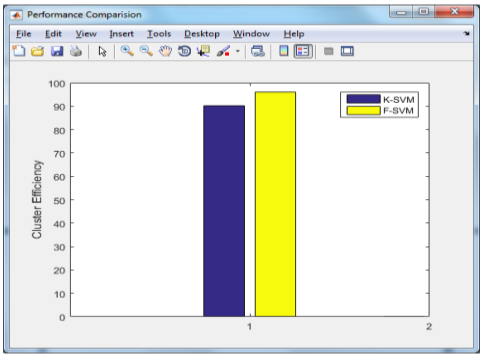

Figure 5: Cluster Efficiency of K-SVM and F-SVM Techniques

V. CONCLUSION

In medical industry there is a huge amount of patient's data for analysis. This healthcare data can be used to extract knowledge for further disease prediction. Currently data mining techniques are widely used in clinical expert systems for prediction of various diseases. Data clustering is a common technique for data analysis and is used in many fields, including data mining, pattern recognition and image analysis. In this thesis, Fuzzy c-means clustering approach is combined with the popular SVM technique (F-SVM), and it is compared with K-SVM algorithm. The result states that the proposed F-SVM algorithm is shown the better performance compared to K-SVM algorithm, in terms of time consumption and cluster efficiency. Therefore, F-SVM clustering approach is useful to explore the malignant and benign structure of breast cancer data.

VI. Future Works

There are various optimization techniques available to achieve high clustering accuracy. Therefore, the future research can focus on improving the accuracy of clusters through optimization technique. This proposed approach is certainly useful in the breast cancer context, whereas it should evaluate with different datasets to confirm the stability of the proposed algorithm. This work can be extended with the inclusion of more dataset and then sampling is enhanced by incorporating distance metric.

VII. REFERENCES

[1]. Al Shalabi, L and Shaaban, Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. In International Conference on Dependability of Computer Systems, DepCos-RELCOMEX'06, pp. 207-214.

[2]. Alba, E, Garcia-Nieto, J, Jourdan, L and Talbi, E.G. (2007). Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In IEEE Congress on Evolutionary Computation, pp. 284-290

[3]. Alshamlan, H.M, Badr, G.H and Alohali, Y.A, (2015). Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational biology and chemistry, vol. 56, pp. 49-60.

[4]. Amini, A, Wah, T.Y, Saybani, M.R and Yazdi, S.R.A.S, (2011). A study of density-grid based clustering algorithms on data streams. In Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol. 3, pp. 1652-1656.

[5]. Amiri, B, Hossain, L, Crawford, J.W and Wigand, R.T. (2013). Community detection in complex networks: Multi–objective enhanced firefly algorithm. Knowledge-Based Systems, vol. 46, pp. 1-11.

[6]. Anderson, P.E., Reo, N.V., DelRaso, N.J., Doom, T.E and Raymer, M.L. (2008). Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. Metabolomics, vol. 4, no. 3, pp. 261-272.

[7]. Baskar, S.S., Arockiam, L and Charles, S. (2013). A systematic approach on data pre-processing in data mining. Compusoft, vol. 2, no. 11, pp. 335.

[8]. Bellazzi, R and Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, vol. 77, no. 2, pp. 81-97.

[9]. Ben-Dor, A, Chor, B, Karp, R and Yakhini, Z, (2003). Discovering local structure in gene expression data: the order-preserving sub matrix problem. Journal of computational biology, vol. 10, no. 3-4, pp. 373-384.

[10]. Blekas, K, Galatsanos, N.P, Likas, A and Lagaris, I.E, (2005). Mixture model analysis of DNA microarray images. IEEE Transactions on Medical Imaging, vol. 24, no. 7, pp. 901-909.

[11]. Boeringer, D.W and Werner, D.H. (2004). Particle swarm optimization versus genetic algorithms for phased array synthesis. IEEE Transactions on antennas and propagation, vol. 52, vol. 3, pp. 771-779.

[12]. Chakraborty, G and Chakraborty, B, (2013). Multi-objective optimization using Pareto GA for gene-selection from microarray data for disease classification. Proceedings of IEEE international conference on systems, man, and cybernetics (SMC), pp. 2629-2634.

[13]. Chang, D.X, Zhang, X.D and Zheng, C.W, (2009). A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recognition, vol. 42, no. 7, pp. 1210-1222.

[14]. Chen, M.S, Han, J and Yu, P.S, (1996). Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and data Engineering, vol. 8, no. 6, pp. 866-883.

[15]. Chen, M.S, Park, J.S and Yu, P.S, (1998). Efficient data mining for path traversal patterns. IEEE Transactions on knowledge and data engineering, vol. 10, no. 2, pp. 209-221.

[16]. Chu, F and Wang, L, (2005). Applications of support vector machines to cancer classification with microarray data. International journal of neural systems, vol. 15, no. 06, pp. 475-484.

[17]. Chu, F and Wang, L, (2006). Applying rbf neural networks to cancer classification based on gene expressions. In International Joint Conference on Neural Networks, pp. 1930-1934.

[18]. Chuang, H.Y, Liu, H, Brown, S, McMunn-Coffran, C, Kao, C.Y and Hsu, D.F. (2004). Identifying significant genes from microarray data. In Fourth IEEE Symposium on Bioinformatics and Bioengineering, pp. 358-365.

[19]. Corso, J.J, Sharon, E, Dube, S, El-Saden, S, Sinha, U and Yuille, A, (2008). Efficient multilevel brain tumor segmentation with integrated bayesian model classification. IEEE transactions on medical imaging, vol.27, no.5, pp. 629-640.

[20]. Crespo, F and Weber, R. (2005). A methodology for dynamic data mining based on fuzzy clustering. Fuzzy Sets and Systems, vol. 150, no. 2, pp. 267-284.

[21]. Damayanti, A and Pratiwi, A.B. (2016). Epilepsy detection on EEG data using back propagation, firefly algorithm and simulated annealing. In International Conference on Science and Technology-Computer (ICST), pp. 167-171.

[22]. A. Bonnaccorsi, “On the Relationship between Firm Size and Export Intensity,” Journal of International Business Studies, XXIII (4), pp. 605-635, 1992. (journal style)

[23]. R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)

[24]. M. Clerc, “The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization,” In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)

[25]. H.H. Crokell, “Specialization and International Competitiveness,” in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)

[26]. K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, “A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II,” KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

[27]. J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/ news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)