HYBRID SUPPORT VECTOR MACHINE BASED MARKOV CLUSTERING FOR TUMOR DETECTION FROM BIO-MOLECULAR DATA

S. SubashChandraBose[1] and T. Christopher[2]
[1]Department of Computer Science, PG and Research Department, Government Arts College, Udumalpet, Tamil Nadu, India
[2]Department of Computer Science, PG and Research Department, Government Arts College, Coimbatore, India
E-Mail: bose.milestone@gmail.com

ABSTRACT

Tumor clustering from gene expression data has paramount implications for cancer diagnosis and treatment. The adoption of clustering techniques for bio-molecular data provides new way for cancer diagnosis and treatment. In order to perform successful cancer diagnosis and treatment, cancer class discovery using bio-molecular data is considered to be one of the most important tasks. Several single clustering approaches were performed for tumor clustering but it had several drawbacks such as stability, accuracy and robustness. In this paper to improve the tumor clustering, we employ a framework, called, Hybrid Support Vector Machine (HSVM) which incorporates PSO-based feature extraction and GA-based feature selection. Specifically, the framework represents the generation of cluster in the first stage which is performed through Markov clustering algorithm. Then, the SVM classification process is adopted to generate or classify the bio-molecular data into benign tumor or malignant tumor. Our experimental results on real datasets collected from UCI machine learning repository and cancer gene expression profile show HSVM can improve the accuracy of clustering gene expression data than other related technique. The Markov clustering algorithm employed in HSVM achieves comparatively better diagnostic performance, capable of classifying the bio-molecular data into benign tumor or malignant tumor based on gene expression data.

Keywords: tumor clustering, hybrid support vector machine, feature extraction, feature selection, Markov clustering, Bio-molecular data.

1. INTRODUCTION

Rapid development of multi clustering techniques for bio-molecular data, more and more researchers are making use of these new techniques,

Which represent in a more precise manner and reliable methods for cancer diagnosis and treatment compared with conventional cancer diagnosis methods based on the single clustering techniques? Multi clustering techniques allow the monitoring of the expression levels of several genes [1] which make classification more feasible with less classification [2] error.

The adoption of tumor clustering approaches to perform cancer class discovery from bio-molecular data provided a new technique for cancer diagnosis and treatment. Standard clustering methods, such as K-means [3], fuzzy C-means (FCM) [4], hierarchical methods, self organizing maps (SOM) [5], simulated annealing based approach [6] and [7] and genetic algorithm (GA) based clustering methods [8] [9] etc. have been utilized for clustering gene expression data.

In order to further improve the performance of tumor clustering from bio-molecular data, in this paper, the HSVM framework integrate five different phases such as Feature extraction, Feature selection, Markov Clustering, Kernel function and classification. The first phase involves the extraction of features that is performed using Principal Component Analysis algorithm. With the extracted features, to reduce the redundancy in the features being selected, the HSVM framework selects the features using Genetic Algorithm.

To the selected features, Markov Clustering (MCL) is applied to cluster together with the proteins having similar biological functions, providing good clustering results and robust against noise in graph data. To find out the cluster center within a group of samples, the proposed HSVM framework employs the Kernel function that groups the nearest group into cluster together. Finally, with the application of SVM, the bio-molecular data is predicted as either benign tumor or malignant tumor.

The experimental results show that the HSVM framework outperforms these comparing algorithms and HSVM framework can serve as an effective technique for bio-molecular data analysis. The rest of this paper is structured as follows. Section 2 provides a review of related works. Section 3 introduces the Hybrid Support Vector Machine framework for cancer bio-molecular data. Section 4 evaluates the performance of our proposed framework using different cancer datasets followed by which discussions are included in Section 5. Section 6 concludes the paper.

2. RELATED WORKS

Several traditional supervised clustering algorithms were initially employed to cluster cancer gene expression data. Modified double selection based semi supervised clustering ensemble framework was presented in [10] performing tumor clustering based on bio molecular data. This method was based on cluster ensemble framework that removed noisy genes present in data by feature selection, adoption of multiple feature selection method for predicting the clustering performance for ensemble approach and finally improved the performance by subset of clustering.

Normalized Expectation-Maximization (EM) algorithm for tumor clustering using gene expression data was implemented in [11]. It was considered to be the first mixture model clustering algorithm that provided stability when clustering with large number of dataset. Hybrid fuzzy cluster ensemble framework (HFCEF) for tumor clustering from cancer gene expression data was implemented in [12]. This framework was the combination of soft clustering and hard clustering into clustering ensemble framework. Two main processes were involved in the design of HFCEF. The first one was the creation of set of new datasets through affinity propagation algorithm and the second one included a consensus function to generate fuzzy matrices and obtain the result as tumor data or non-tumor data.

Though classification of cancer has improved over the last two decades, prediction by gene expression monitoring was still considered to be in the preliminary stage. Cancer class discovery and prediction using Neighbourhood analysis was presented in [13]. Yet another work of classification of multiclass cancer diagnosis using tumor gene expression was investigated in [14] [15] resulting in the improvement of classification analysis.

In this paper, we investigate the hybrid SVM and study its performance in clustering cancer gene bio-molecular data. Particularly, HSVM framework leverages the gene-to-cluster to disclose the underlying pattern of cancer gene bio-molecular data. In addition, HSVM integrate the advantages of high quality data being extracted and grouping the nearest group into cluster using Markov Clustering algorithm to mitigate the intrinsic issues (i.e., high dimensionality, few samples, and many noisy genes) of clustering gene bio-molecular data. Our experiments on various publicly available cancer gene expression data demonstrate that HSVM group samples more accurately than the state-of-the art methods discussed.

3. HYBRID SUPPORT VECTOR MACHINE

In recent year, different clustering approaches are adapted by several researches for tumor clustering from bio-molecular data. Compared with other single clustering algorithm, the proposed Hybrid Support Vector Machine (HSVM) framework adopted multi cluster model with which the clustered results are used for classification. Figure-1 shows the block diagram of HSVM framework.

This HSVM framework comprises of two major steps, namely clustering and classification. The objective of the first step is to generate a set of clustering. The second step focused on classifying an individual set of clustering and with which the results are predicted to be either benign tumor or malignant tumor through SVM classification.

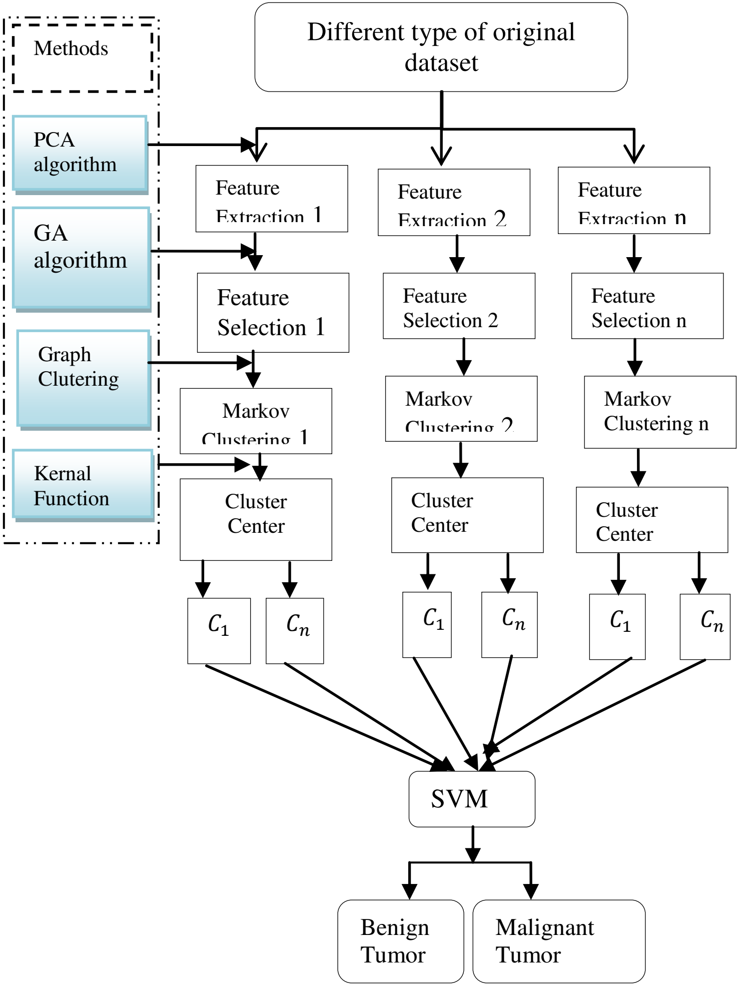

Figure-1. Block diagram for hybrid support vector machine.

The predicted results through HSVM framework were found to be more accurate, stable and robust.

3.1 Principal component analysis-based feature extraction

Principal component analysis [16] is standard method used for statistical pattern recognition and signal processing for data reduction and feature extraction [17]. Principal Component Analysis serves as the basic feature extraction model in our work, which is thus referred to as Principal Component Analysis-based Feature Extraction (PCA-FE). Let us consider an input tumor image of size $T \times T$ pixels, represented by one dimensional vector 'T^2'.

The main goal of PCA algorithm remains in finding the vectors for the distribution of images within the entire image space. Let us assume the training set of input tumor images being represented as '$I_1, I_2, .., I_N$'. Then, the average value 'Ψ' for above training set is mathematically formulated as given below.

$$\Psi = \frac{1}{N}\sum_{t=1}^{N} I_t \tag{1}$$

Also, the resultant average value is different for each tumor image 'Φ_i' present in the given database, and is mathematically given as below.

$$\Phi_i = I_i - \Psi \tag{2}$$

PCA is subjected by highly dimensional vector that in turn reduces the amount of redundant information through de-correlation of input vectors. Through de-correlation of input vectors, the PCA-FE seeks a set of 'n' ortho normal vector represented by 'Ortho_n', to determine the distribution of the data and is represented as given below.

$$\lambda_r = \frac{1}{N}\sum_{m=1}^{N} (O_r^T \Phi_m)^2 \tag{3}$$

The objective of deriving the above mathematical formulation is to maximize the variance in the direction of principal vectors subjected to as given below.

$$\mathrm{Ortho}_I^m \mathrm{Ortho}_r = \delta_{ir} = \begin{Bmatrix} 1, \text{if } I = r \\ 0, \text{otherwise} \end{Bmatrix} \tag{4}$$

Now the solution to the maximization problem is to evaluate the eigenvectors [18] and the eigen values of the covariance matrix is mathematically formulated as given below.

$$CM = \frac{1}{N}\sum_{t=1}^{N} \Phi_t \Phi_t^T = AA^T \tag{5}$$

From above equation, matrix, '$A = [\Phi_1 \Phi_2 \dots \Phi_N]$', let us consider eigenvector 'e_i' of '$A^T A$' such that

$$A^T A e_i = \mu_i e_i \tag{6}$$

The above equation is multiplied by matrix A is then mathematically formulated as given below.

$$AA^T {Ae_i}^T = \mu_i A e_i \tag{7}$$

From above equation, 'Ae_i' defines the eigenvector 'e' whereas 'μ_i'eigenvalue is defined by '$C = AA^T$'. With these analysis, a matrix '$M \times M$' is constructed and 'M' eigenvectors, 'e_i' of 'L' is measured. The above vector defines the linear combination of M training set images to form the eigenvalues 'O_I' and is mathematically formulated as given below.

$$O_I = \sum_{k=1}^{H} e_{ik} \Phi_k, I = 1, .., H \tag{8}$$

With this analysis, the calculations are greatly reduced, from the order of the number of pixels in the images T^2 to the order of the number of images in the training set (H). As a result, PCA-FE provides high quality features for tumor image recognition

3.2 Genetic algorithm-based feature selection

With the extracted features, each feature in the principal component is considered as a feature vector and Genetic Algorithm (GA) is used for feature selection. Genetic algorithm is considered to be one of the search-based optimization methods that entire depend on the principle of natural selection and genetics [19] [20] [21]. A general flow diagram of genetic algorithm is shown in Figure-2

GA identifies an optimal feature subset. In feature selection problem, each individual feature subset encodes the decision variables of search space into finite length where the quality of each feature is evaluated with the aid of a fitness function.

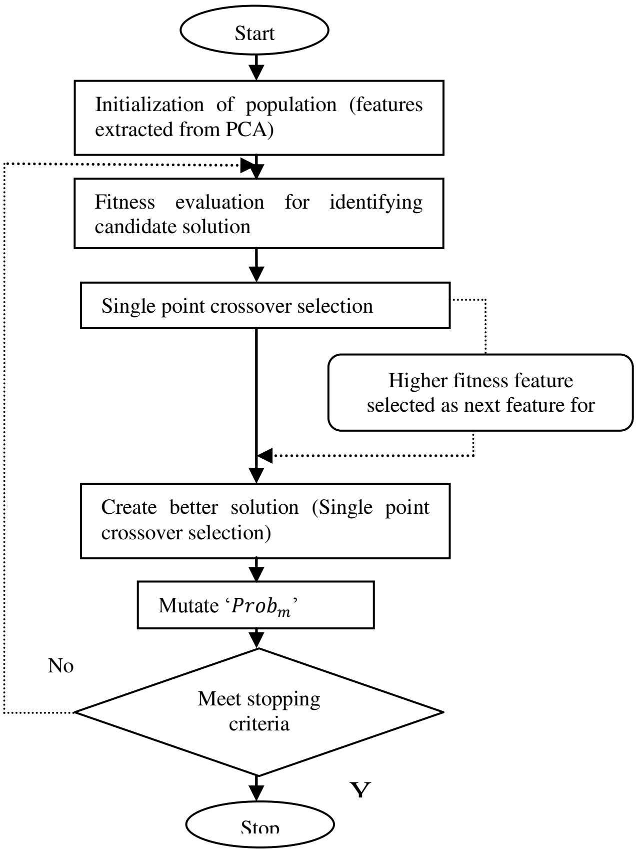
Start
Initialization of population (features extracted from PCA)
Fitness evaluation for identifying candidate solution
Single point crossover selection
Higher fitness feature selected as next feature for
Create better solution (Single point crossover selection)
Mutate '$Prob_m$'
No
Meet stopping criteria
Y
Stop

Figure-2. Genetic algorithm-based feature selection.

A good distinguish between good and bad solution is then made implementing the natural solution. The following seven steps are used to select the extracted features for search problem.

The first step involved in GA-FS is the initialization where the candidate solution initial population is generated randomly across the search space. Here each chromosome in the overall population represents the candidate solution across the search space. Once the initialization of population is performed, the fitness value is evaluated for identifying the candidate solution, where the propose HSVM uses Fisher Criterion that represents the ratio of between-class scatter to the within-class scatter.

The next step in the proposed GA-FS is represented by the selection process that selects the best solution to worse ones. The solution that has higher fitness value is then assigned as the next feature for later breeding. Followed by this, single-point crossover operator is selected that creates new better solution i.e., offspring based on the combination of two or more parental solution. Finally, each individual generated through single-point crossover has a probability of '$Prob_m$' to mutate. The above steps are repeated until relevant features are obtained. Algorithm 1 given below provides an overview of the GA-FS for obtaining the relevant features.

```
Input: population, features extracted
Output: Most relevant features selected
Initialize population with random candidate solutions
1: Begin
2: Evaluate each candidate solution
3: Repeat
4: While termination condition is not true do
5: Select individuals for the next generation
6: Recombine pairs of parents
7: Mutate the resulting offspring
8: Evaluate each candidate solution
9: End while
10: Until relevant features are obtained
11: End
```

Algorithm 1 Genetic algorithm-based feature selection

As given above, the GA-FS performs three major steps. For each candidate solution (i.e. features extraction through PCA), the GA-FS selects individual features for the next generation. Followed by this, fitness function is evaluated to either to consider the feature to be used for next breeding or consider the same feature generated through single-point crossover operator. Followed by this, the resultant offspring is mutated and the process is repeated until relevant features are obtained

3.3 Markov clustering algorithm

With the relevant features selected using GA, the proposed HSVM framework applies a Markov Clustering (MCL) algorithm based on stochastic flow simulation that provides an effective result in clustering biological networks.MCL algorithm provides several advantages based on transition in graphs [22]. The Markov clustering algorithm (MCL) simulates random walks as shown in Figure-3 and its corresponding Markov Chain Cluster is shown in figure 4 on the underlying interaction network, by alternating two operations: expansion, and inflation.

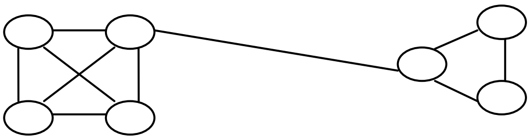

Figure-3. Random walks.

The first operator expansion in MCL is responsible for connection between different regions (i.e. different features) of the graph. Finally, this graph is translated into a Markov matrix. The edge weights are higher in links that are found to be within the cluster whereas lower in links between the clusters. For each vertex the transition values (i.e. the selected features) of MCL in HSVM framework strengthens strong neighbours whereas the less popular neighbours are weakened.

The two operations are repeated until the graph separate into two subsets. Finally there is no longer path between theses type of subsets.

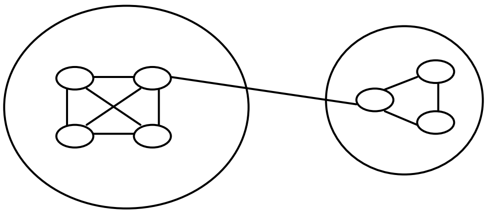

Figure-4. Markov chain cluster structure.

Markov chain cluster structure

The cluster center or exemplars is identified by Affinity Propagation (AP). Initially, all nodes in graph all nodes are considered as cluster center, behind this there is no prior knowledge for the selection of cluster center in graph. AP computes the responsibility '$r(i,k)$' for each node 'i' and each candidate exemplar 'k', this responsibility defines how well 'k' is an exemplar for node 'i', and this is reflected by the availability $a(i,k)$ that I should choose 'k' as an exemplar. This is mathematically formulated as given below.

$$r(i,k) \leftarrow s(i,k) - \max_{k'k' \neq k}\{a(i,k') + s(i,k')\} \tag{9}$$

$$a(i,k) \leftarrow min\{0, r(k,k) + \sum_{i':i' \in \{i,k\}} max\{o, r(i',k)\}\} \tag{10}$$

From (9) and (10), '$s(i,k')$', defines similarity between the two nodes (i.e. features) 'i' and 'k'. The above steps are repeated until a steady state is converged. The resulting matrix thus arrived to discover clusters.

3.4 Kernel function

The output of MCL provides a set of clusters. The Kernel function finds out the cluster center within the set of cluster. Let us consider the set of cluster as '$S_1, S_2, \ldots, S_n$', where 'S_1' represent the set of cluster from one type of database, 'S_2' represents the set of cluster from another input database and so on. The Kernal function finds the cluster center for each set presented in different databases using different databases given below.

$$K_1(x, x')exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \tag{11}$$

From (11), 'x' represents one group of cluster in set of cluster present in database x'represents another group of cluster present in set of cluster in same database. The above equation is used find out the cluster center for other set of cluster present in database and so on. Through cluster center '$K_1, K_2, .., K_n$' from each database it is easy to find out the nearest cluster and it is grouped together in to two samples such as:

$$K_1 = C_1 \ and \ C_2 \tag{12}$$

$$K_2 = C_1 \ and \ C_2 \tag{13}$$

$$K_n = C_1 and \ C_2 \tag{14}$$

From (12), (13) and (14), '$K_1, K_2, .., K_n$' are cluster center and '$C_1 and \ C_2$' represents the nearest clusters that are grouped together into two groups.

3.5 SVM-based classification

Finally, the HSVM framework using SVM, supervised learning method to predict [23] [24] output as either benign tumor or malignant tumor, with the advantage of the method being used as a wide range for pattern recognition problem. The binary classification (i.e. benign tumor or malignant tumor) in the proposed HSVM framework is constructed using hyperplane that separates class members from non-members in the input space.

Let us consider a training example '(x_i, y_i)', where x_i represents the real data instance and 'y_i' indicates the labels that belongs to the class instance. Binary classification in HSVM framework includes two classes for recognition of benign tumor or malignant tumor i.e., '$y_i = +1$' or '$y_i = -1$'.

If the training set (x_i, y_i) is positive then assigned $y_i = +1$ and predicted as malignant tumor otherwise assigned negative to the class label and predicted as benign tumor. The objective behind the use of SVM is the construction of a hyperplane for achieving maximum separation between two classes. Separating the classes with a large margin minimizes a bound on the expected generalization error, with the existence of vector 'W' and a scalar 'b' and is represented as given below.

$$y_i(W.\mathrm{x}_i + b) - 1 \geq 0 \tag{15}$$

Using set of function hypothesis space is defined mathematically as given below.

$$f_{w,b} = sign(W.X + b) \tag{16}$$

Separating hyperplanes for which the distance between the classes is identified by SVM classifier along a linear perpendicular to the hyperplane is maximized.

This is obtained by solving the below optimization problem.

$$Minimize \ \frac{1}{2}\|W\|^2 \tag{17}$$

In linear cases, there is a restriction that given class in the training case all lie on both the side of the hyperplane and is said to be relaxed using a slack variable such as:$\xi_i \geq o$. In this situation there is occurs two condition, where SVM searches the hyperplane used for the maximization of the margin and second one is minimization of the misclassification error. If there arises any trade off between these misclassification error and margin, the HSVM framework controls using a positive constant 'C'. For non-separable data the optimization problem is re-written as given below.

$$\min_{w,b,\xi_1,..,\xi_k}\left[\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{k}\xi_i\right] \tag{18}$$

Hence, with the above optimized results, the bio-molecular data, diagnosis of cancer is made in an efficient manner. Through this, the results are classified as either benign tumor or malignant tumor minimizing the classification error.

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed framework for applying multi cluster model to bio-molecular data for cancer diagnosis and classification, comparative experiment was conducted on benchmark datasets. The proposed tumor clustering from bio molecular data are evaluated by the datasets from cancer gene expression profiles namely St. Jude leukemia [1], Lung cancer (LC) [2] respectively

Table-1. Cancer gene expression datasets.

Dataset	Source	Samples
Leukaemia	Cancer gene expression	72
Lung cancer	Cancer gene expression	32

The two datasets Leukemia and Lung cancer includes, of 72 and 32 Samples. For Performance evaluation two datasets are considered such as Lung cancer and Leukemia.

The performance is evaluated for HFC and HSVM classification methods in terms of Accuracy, sensitivity, specifity, precision, recall and Fscore. Table-2 tabulates the performance evaluation of HFC and HSVM using Leukemia dataset.

Table-2. Performance evaluation of HFC and HSVM using Leukemia dataset.

Metrics	HFC	HSVM
Accuracy (%)	57.5059	60.3333
Sensitivity (%)	59.8587	62.5263
Specificity (%)	55.2536	57.8824
Precision (%)	60.4158	62.5263
Recall (%)	59.6930	62.5263
Fscore (%)	59.8675	62.5263

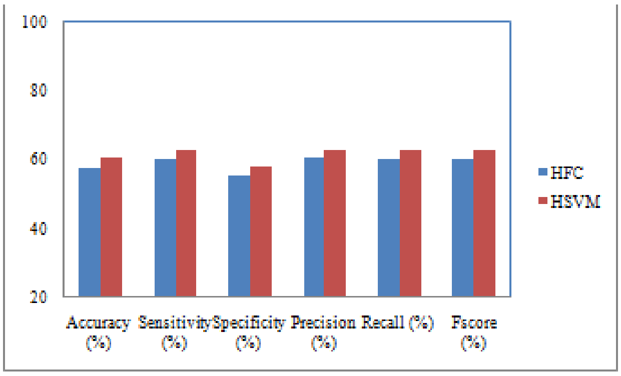

Figure-5. Comparative results of HFC and HSVM using Leukemia data.

Figure-5 shows the comparative results of HFC and HSVM using Leukemia dataset. As shown in the Figure-5, six metrics accuracy, specificity, sensitivity, precision, recall and fscore used for analyses outperforms the existing HFC. The improvement observed in HSVM is due to the extraction of features using principal component analysis that extracts features with high by eigen value decomposition for each attribute. As a result, bio-molecular data with high dimensional data space extracts the first few principal components with the objective of reducing the dimensionality of the transformed data. The cluster generated for Leukemia dataset is shown in the Figure-6.

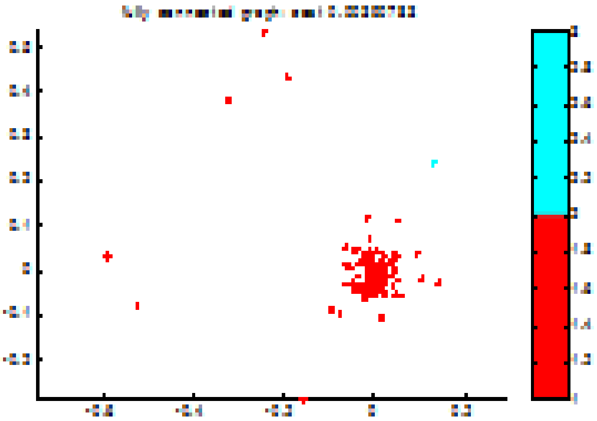

Figure-6. Cluster generation for Leukemia dataset.

As shown in the Figure-6, minimum number of related clusters is generated using GA. The features selected using GA is not confined to single solution, but repeatedly modifies the population of bio-molecular data. The population size in HSVM is selected in a random manner that further in turn allows in identifying the entire range of population. This in turn helps in the improvement of HSVM framework (Table-2) with 60 % accuracy, 62 % sensitivity, 58 % specificity, 63 % precision, 63 % recall and 63 % Fscore which is greater than existing HFC in all aspects.

Table-3. Performance evaluation of lung cancer dataset.

Metrics	HFC	HSVM
Accuracy (%)	72.6788	74.9064
Sensitivity (%)	92.0207	94.0863
Specificity (%)	30.6215	33.2500
Precision (%)	73.5618	76.4186
Recall (%)	91.2350	94.0863
Fscore (%)	81.8490	84.3151

Table-3 given above shows the tabulated results for six chosen metrics, accuracy, sensitivity, specificity, precision, recall and fscore respectively. From the table we can infer that the performance improvement of the proposed HSVM when compared to HSC in terms of all metrics.

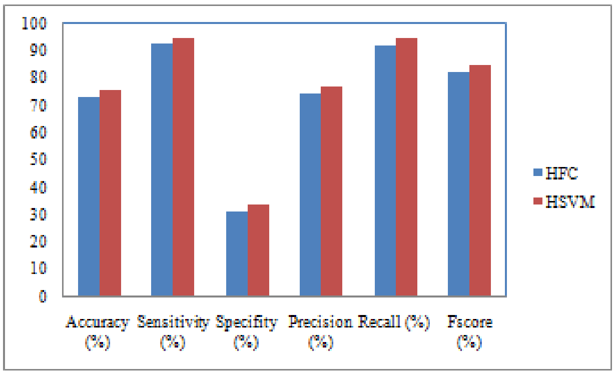

Figure-7. Comparative results of HFC and HSVM.

The above Figure-7 shows the performance comparison of HFC and HSVM classification for Lung cancer dataset. Performance improvement is observed with respect to six metrics (i.e. accuracy, sensitivity, specificity, precision, recall and fscore) using the proposed HSVM framework than when compared to HFC. This performance improvement is due to the de-correlation of input vectors resulting in ortho normal vector. As a result, PCA-FE produces high quality features. With this high quality features being extracted, employing Fisher Criterion in GA-FS the ratio of between-class scatter to within-class scatter is reduced. Hence, we observe performance improvement in the proposed HSVM when compared to HSC. Figure-8 given below shows the cluster generation for Lung Cancer dataset.

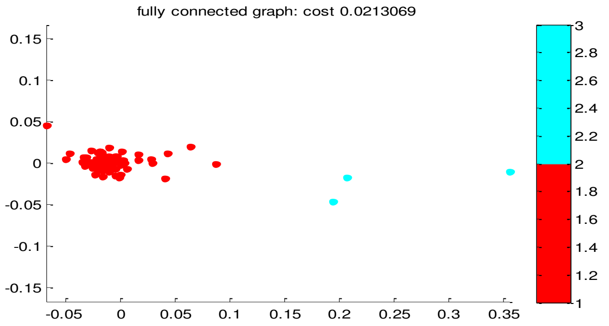

Figure-8. Cluster generation for lung cancer dataset.

With higher relevancy rate being obtained using GA-FE in HSVM by applying Markov Clustering algorithm, strong neighbours are selected i.e. more prominent features related to each other are used as the feature for breeding than the less known features. This is a way through which Affinity Propagation is arrived at discovering clusters. The proposed HSVM achieves (Table-3) 75 % accuracy, 94 % sensitivity, 33 % specificity, 76 % precision, 94 % recall and 84 % Fscore that is greater than existing HFC method in all aspects.
Purity $PU(P,P')$is calculated using,

$$PU(P,P') = \frac{1}{n}\sum_{i=1}^{k} \max_{j\epsilon\{1,\dots,k'\}}\left|C_i \cap C_j'\right| \tag{19}$$

Purity value ranges from 0 to 1. For high PU value, corresponding satisfactory clustering is achieved. The comparison of Purity and Normalized Mutual Information for HFC and HSVM classification is shown in the Figure-9 given below.

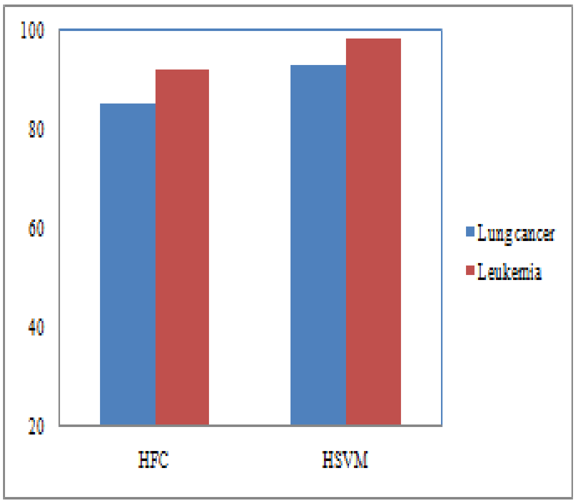

Figure-9. Comparative result of purity using HFC and HSVM.

The above Figure-9 shows that the proposed HSVM method achieves high purity values for Lung cancer, Leukemia datasets. For every regarded case the results of the proposed HSVM framework were compared with the results provided by HFC. However, among the regarded methods, the proposed HSVM framework showed better sign of improvement. The comparison was performed in terms of the purity or the obtained cluster quality that is measured in terms of percentage (%). Based on the graphical results, it can be seen that applying the Markov algorithm for clustering gene bio-molecular data, significantly improves the purity. For the regarded cases using Lung cancer and Leukemia datasets, the average purity was improved by 93% using Lung cancer dataset and 97% using Leukemia dataset respectively.

The proposed method achieves reasonable improvement than existing method. The improvement achieved is due to the construction of hyperplane that in turn helps in achieving maximum separation between two classes (i.e. benign tumor or malignant tumor). Based on data shown in Table it can be also seen, that for most of the cases using hyperplane increased accuracy of image classification. The accuracy of image classification was also increased by using binary classification. This however paid for the significant increase of the specificity and precision.

5. CONCLUSIONS

In this paper we present a new framework for tumor clustering from cancer gene expression profiles. The proposed framework estimates the principal components using PCA and eigen vectors. In other words, this method estimates the most probable sequence of principal components through de-correlation of input vectors instead of making a hard decision on which components to use for the feature extraction. This way, the proposed framework obtains the most high quality features and therefore obtains better results in terms of accuracy compared to some well-known algorithms in this field. Furthermore, the algorithm GA-FS has a lower time-complexity where using the Fisher criterion selects only the relevant features compared to some benchmark methods in terms of sensitivity and specificity. Finally, compared with existing clustering method, the proposed work highly identifies cancer samples of different types. Through the experiments it is observed that proposed method achieves high performance for all datasets. In future, methods to improve the efficiency of HSVM will be implemented and theoretical analysis of HSVM will be performed. The performance of proposed framework will be explored with some other metrics such as disassociation measure and squared error distortion measure.

REFERENCES

[1] E. J. Yeoh, M. E. Ross, *et al.* 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell. 1: 133-143.

[2] A. Bhattacharjee, W.G. Richards, J. Staunton, *et al.* 2001. Classification of human lung carcinomas by mRNA expression profiling reveal distinctadenocarcinomas sub-classes. In Proceedings of the National Academy of Sciences. 98(24): 13790-13795.

[3] R. Herwig, A. Poustka, C. Meuller, H. Lehrach, J. O'Brien. 1999. Large-scale clustering of cDNA fingerprinting data, Genome Research. 9(11): 1093-1105.

[4] D. Dembele, P. Kastner. 2003. Fuzzy c-means method for clustering microarray data, Bioinformatics. 19(8): 973-980.

[5] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proceedings of the National Academy of Sciences. 96: 2907-2912.

[6] U. Alon, N. Barkai, D.A. Notterman, K. Gishdagger, S. Ybarradagger, D. Mackdagger, A.J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences. 96: 6745-6750.

[7] A.V. Lukashin, R. Fuchs. 2001. Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters, Bioinformatics. 17(5): 405-414.

[8] S. Bandyopadhyay, A. Mukhopadhyay, U. Maulik. 2007. An improved algorithm for clustering gene expression data, Bioinformatics. 23(21): 2859-2865.

[9] U. Maulik, A. Mukhopadhyay, S. Bandyopadhyay. 2009. Combining Pareto-optimal clusters using supervised learning for identifying co-expressed genes, BMC Bioinformatics. 10(27).

[10] Zhiwen Yu, Hongsheng Chen, Jane You, Hau-San Wong, Jiming Liu, Le Li and Guoqiang Han. 2014. Double Selection Based Semi-Supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 11(4).

[11] Nguyen Minh Phuong; Nguyen Xuan Vinh. 2008. Normalized EM algorithm for tumor clustering using gene expression data. in Bio Informatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on. pp. 1-7, 8-10.

[12] Zhiwen Yu; Jane You; Hantao Chen; Le Li; Xiaowei Wang. 2012. Tumor clustering based on hybrid cluster ensemble framework. in Computerized Healthcare (ICCH), 2012 International Conference on. pp. 95-101.

[13] T. R. Golub, D. K. Slonim, P. Tamayo, *et al.* 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression. Science. 286: 5439, pp. 531-537.

[14] S. Ramaswamy, P. Tamayo, R. Rifkin, *et al.* 2001. Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures. Proceedings of the National Academy of Sciences. 98(26): 15149-15154.

[15] A.I. Su, M. P. Cooke, *et al.* 2002. Large-scale analysis of the human andmouse transcriptomes. Proceedings of the National Academy of Sciences. 99(7): 4465-4470.

[16] Gumus, E., Kilic, N., Sertbas, A., & Ucan, O. N. 2010. Evaluation of face recognition technique using PCA, wavelets and SVM. Expert Systems with Applications. 37, 6404-6408.

[17] Haykin S. 1999. Neural Networks: a comprehensive foundation, (2nd ed.). Prentice-Hall.

[18] Turk M. & Pentland A. 1991. Eigenfaces for recognition. Journal of Cognitive Neuroscience. 3(1): 71-86.

[19] Fraser A. S. 1957. Simulation of genetic systems by automatic digital computers. II: Effects of linkage on rates under selection, Austral. J. Biol. Sci. 10: 492-499.

[20] Bremermann H. J. 1958. The evolution of intelligence. The nervous system as a model of its environment, Technical Report No. 1, Department of Mathematics, University of Washington, Seattle, WA.

[21] Holland J. H. 1975. Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI.

[22] J. Vlasblom, S.J. Wodak. 2009. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs, BMC Bioinform. 10: 99.

[23] V. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer, N.Y. ISBN 0-387-94559-8.

[24] Burges C. 1998. A tutorial on support vector machines for pattern recognition. In Data Mining and Knowledge Discovery. Kluwer Academic Publishers, Boston. Vol. 2.