DOI: 10.21917/ijsc.2021.0349

A HYBRID ENSEMBLE METHOD FOR ACCURATE FUZZY AND SUPPORT VECTOR MACHINE FOR GENE EXPRESSION IN DATA MINING

S. Vasanthakumar and N. Ranjith

Department of Computer Science, KSG College of Arts and Science, India

Abstract

Malignancy is a bunch of infection which spreads all through the human body. Since it is an exceptionally deceptive illness its determination is of vital importance. Information mining innovation helps in arranging and bunching the malignancy information and this procedure assists with distinguishing potential disease patients by investigating the data alone. In this examination we analyze three information mining calculations, namely PCA, Genetic calculation and Hierarchical Fuzzy C Means (HFCM). The hereditary calculation is done using the Quantum-enhanced Support Vector Machine (QSVM). The outcome demonstrates that the proposed calculation accomplishes a better outcome when contrasted to the other two calculations.

Keywords:

PCA, Genetic Algorithm, Hierarchical Fuzzy C Mean, QSVMs, Cluster

1. INTRODUCTION

This study aims at introducing a framework for determination and prediction of malignant growth illness using data mining procedures. The information mining is calculated using crossover technique under classification and clustering in this framework. Inference of disease is significant as the discovery of malignant growth in the initial phase can help in giving appropriate therapy to control the growth of the malignant tumor. Hence this framework is useful in studying the growth of cancer.

The importance in clinical field is the early identification of any illnesses which assist in treating it in the initial phase and preventing the future growth of the disease. Malignancy is one such disease where the early identification can decrease the death rate in the disease affected patients. There are by and large two kinds of malignancy disease, 1) Benign disease and 2) Malignant Cancer.

The Benign disease grows the conditions of the tumor for the non-cancerous it cannot be spreads to the other parts of the cells. Sometimes it causes to the dangerous for the diseases itself. The malignant is to spread the other parts of the cells in the living organisms is spreads faster and easier for the other parts faster.

There are chances that the malignant growth may be identified in the initial stage preventing the increase of further patients. Colon and cell splitting in the lungs is one of the main diseases in many countries including India.

It is the second major reason for deaths caused by diseases amongst the people. The dangerous tumor is created when the cells in the tissues get partitioned and develop without the ordinary controls of cell growth and separation. Information mining methods can be utilized to foresee the malignancy in a patient utilizing different manifestations and data from the previous outcomes. Significant and relevant data can be identified through these information mining strategies.

2. LITERATURE REVIEW

This work portrays the AI field and Relief calculation that depends on malignancy information collected. The SVM and k-NN is tested and the outcome is compared. The Relief-f with the sifting strategy incorporate addition, acquiring proportion and x^2 measurement and the value of Relief-f has great precision [1].

The computation is referred to as a hereditary weighted k-. A hybridization of hereditary calculation is the genetic weighted k-means algorithm (GWKMA). WKMA is the inferred hereditary choice, hybrid, change activity. The outcome of the computational articulation is that GWKMA performs better than k-implies.[2].

The approach is used to the selection of the arrangement technique with the presentation of direct separate examination and its change strategy for grouping of high-quality articulation data. The approach includes forecasting for tiny clusters, Partition around Medoids (PAM), and shrinking centroid, which consists of shifting the threshold to the center position in this setting, is equivalent to zero on its threshold values. In comparison to the typical outcomes, the outcome is deferential. When it comes to change tactics, there is not much of a difference.

In k-NN calculation for quality articulation information on disease classifications, the important ones are positioned by the presentation utilized by eager methodology in the preparation information. The projected PIC-NN gathering border allows for better displays.[3].

The novel system of edge-based example weighting which generally investigates the realistic examples and furthermore changes an effective calculation beneath the structure utilizing miniature cluster information. The result shows that the example weighting calculation has a regular procedure [4].

DNA microarray techniques to measures the large number data conditions for expression levels of thousands to be calculate the genes under various experimental conditions. After doing the several preprocessing steps and getting the results of lower level microarray analysis data, the microchip can be represented as a digital matrix with rows and columns corresponding to calculate the results and doing the experimental conditions. Gene sequence vectors are called gene expression models, and columnar vectors are called conditional expression profiles. In higher level microarray analysis, the data extraction techniques are used to analyze large amount of data in the biological information [5].

The time arrangement during a biological cycle under the test of numerous information, the diverse microarray procedures on a distance measure with different bunching calculations for evaluating the clustering result [6].

The highlight determination strategies are important to define and calculate the highest point and identify another technique to gauge the estimation of all highlights by applying SNR. The outcome of utilizing techniques can produce great outcome in the exact order of time [7].

The various methods imply the filtering approach for gene expressions to the clustering methods. Different Approaches to Filtering Gene Expression Data for assembling Comparisons. Grouping is the task of classifying groups of people so that people in one set (called a group) are more identical than others. Microchips are clusters of microscopic DNA that adhere to a solid exterior. Microarrays are used to measure the expression and compare the level of many genes sequence. A gene can be collected in a set and this group contains of two or more genes. This resembles the similarity of code for the same or a similar product. Common ancestors have and share the same group of genes, which helps to track the heredity and generations. The proposed work with different filtering approaches such as entropy, Genevar filter and Genelowval filters are compared [8].

Clustered algorithm combines mutual information in genes and coexpression conditions using a single parameter and is recognized to function better than other current algorithms. It is possible to identify several data types categories that are connected to the such as proteome, microarray, metabolomics, and many more [9].

In this progress, the protein data screens information about someone needed to understand different for genetic relationship processes associated with the physical environment. Identifying unknown patterns in protein data offers a great advantage in improving the functions and interactions among the proteins. The state is become more complicated to find the biological tissues and also gene volumes make it difficult to understand and interpret the resulting large amounts of data, consisting of millions of measurements. The data obtained also helps to avoid uncertainty, inaccuracy and disturbances. So, using clustering techniques is the first step towards overcoming this problem. This is important for discovering natural structures during data mining and identifying patterns in underlying data [10].

The classification of rectal cancer is done using selection and group approach to not have any errors and improve the response to treatment in cancer clinical trials. The tumor acts as a group that makes genetic changes and identifies the disorders. One of the techniques to imply the microarray has been broadly developed to measure the changes in gene expression in different levels under normal levels and experimental conditions. Typically, protein data are characterized by different levels of sizes from larger and smaller. Therefore, the selection of datasets is necessary to find the lowest number of informative genes and it helps to improve the classification accuracy and results of genetic relationship interpretation. Similar groups of genes are used because methods adopted for selecting the traits ignore the exchanges between the genes. In addition, the efficiency of the classifier can be determined from the quality of the selected data. This research provides an opportunity to group and select the characteristics [11].

In the fields of bioinformatics and clinical research, microarrays are often used to distinguish between datasets of cancer between without abnormal and displays the tumor samples. More over the larger size of datasets it contains the affects data is more it affects the accuracy of the experimental classification. Therefore, it is necessary to select traits to separate the information about genes and avoid genes that do not have information and are redundant. However, some selection methods avoid the protein interactions it causes the different results. Thus, the similar or common genes are cluster within the one and different types or affected genes are grouped to another using the clustering methods. To find the higher classification accuracy and easily

Analysis of gene expression data is important for gene therapy and cancer diagnosis. Clustering is one of the important methods to analyzing the protein data and executes the better results. The data are often characterized by a large number of genes easily, but as the samples are limited the accuracy becomes higher. Therefore, different estimations and ensemble cluster methods are proposed to solve these problems [13].

The method of classifying genes is a very significant area of research as it is used in prognostic and diagnostic systems for acquired cancer. Genetic evaluation consists of thousands of samples that are taken to produce varying results. Therefore, accurate and efficient methods to evaluate the data samples are become more difficult [14].

3. COMPARISON OF ALGORITHMS

3.1 PCA ALGORITHM

Principal Component Analysis (PCA) is a method to perform linear mapping of the data in lower levels of dimensional space is needed that is often used to makes the smaller dimensionality of huge datasets. In this method can be improves the attained by converting a large volumes of dataset variables into a lower one that still it contains more information in the huge dataset.

It contains the minimum number of variables in the original dataset and helps to improve the accuracy, but while using the linear mapping process, the value becomes smaller and the accuracy rate and simplicity of performing is higher. The reason being that smaller dataset are easier to inquire and visualization. They trend to make easily to analyzing data is become much easier and faster for machine learning algorithms without irrelevant variables to process.

Algorithm for PCA

Step 1: Standardization of the process

Step 2: Joint variability of the random Matrix Computation

Step 3: To calculate the Eigen vector and values of the random Matrix to Identify the Solution of basic Components

Step 4: Vector contains multiple objects

Step 5: Another form of data transformation along with the basic Components methods

3.2 GENETIC ALGORITHM

A Genetic Algorithm (GA) is a method for doing different experimental approach to solve the problems or identify the value by proposing a cognitive style that is inspired by one of the basic the ories said by Charles Darwin. The theory states that the evaluation of different stages occur by the natural selection in different timelines. These algorithms identify and study the individuals in a population is that are different. The fittest that allows an individual to both survive and grow are identified. These individuals are used for the process of reproduction in order to produce the fittest and keep life continuing for the forthcoming generations.

The process of various survival begins with the selection of the appropriate individuals from a diverse collection of people. It is based on the genetic expression is carried out to the next generation procedure to retrieve from the characteristics or characters from the methods of inheriting the properties of the parents. If the parents' generation is more physically fit, they will have a higher chance of surviving obstacles than their parents did. This process repeats itself, step by step, and at the end of it, a generation will be more interconnected and people will be able to grow more readily. The same approaches may be used to solve a search problem and get efficient results. A set of solutions for a problem is considered, and the best techniques among them are chosen.

Algorithm - Genetic Algorithm

Step 1: To executed at the beginning population of the problem

Step 2: Fitness function to be process

Step 3: Selection procedure

Step 4: Crossover techniques

Step 5: Mutation process

3.3 HIERARCHICAL FUZZY C MEANS ALGORITHM

To allocate the number of members in a group to each data point, the Hierarchical Fuzzy C Means (HFCM) method is employed. This approach is used to discover the cluster center and the distance between the cluster center and the data point. If the data is extremely close to the cluster center point, its members are immediately joined to the specific cluster center, causing the values to improve. The iteration procedure can be repeated, with each data point's members being assigned to be equal to one another in the cluster. The cluster centers points grow more moderate throughout the procedure after each iteration.

Algorithm of HFCM

Step 1: Equally allocated the '*c*' cluster point position.

Step 2: Related in-depth of the fuzzy membership μ_{ij} using the calculation:

Step 3: Calculate the fundamentals of fuzzy with center equation.

Step 4: Repeat the steps till the end getting the accurate value is achieved.

3.4 QUANTUM-ENHANCED SUPPORT VECTOR MACHINE (QSVM)

The classification of algorithms and the methods used for machine learning plays an important role in recognizing the pattern and in data mining applications. An important concept in classification methods is that of a kernel. It is not possible to separate out the Data using a hyperplane in its actual space. A common procedure used to identify a hyperplane comprises of applying a non-linear transformation function to the data.

The steps involved for QSVM process are as follows,

First, the classical data point X is translated into a quantum datapoint $|\Phi(\vec{x})\rangle$. This can be achieved by a circuit $V(\Phi(\vec{x}))$. Where $\Phi(\ldots)$ could be any classical function implemented on the classical related data $\vec{x}$.

Secondly, a parameterized quantum circuit $W(\Theta)$ is needed with parameters Θ that progress the data in continuous way that in the end.

In final step involves applying a measurement that returns a classical value -1 or 1 for each classical input $\vec{x}$ that identifies the label of the classical data.

In the QSVM derived the classical functions are defined as

$$\Phi_u(\vec{x}) = x_u \text{ and } \Phi_{uv}(\vec{x}) = (\pi - x_u)(\pi - x_v). \tag{1}$$

The classical data vector $\vec{x} = (x_1, x_2)$ the feature map takes the form:

$$\mathbf{U}(\Phi(\vec{x})) = \exp(i\{x_1\mathbf{Z}_1 + x_2\mathbf{Z}_2 + (\pi - x_1)(\pi - x_2)\,\mathbf{Z}_1\mathbf{Z}_2\}) \tag{2}$$

4. EXPERIMENTAL RESULT

The Colon cancer dataset normally used in earlier studies in gene selection and classification methods. It consists of the gene evaluation profiles of 2,000 genes from 62 tissue samples among which forty are colon abnormal tissues and 22 are normal tissues. The splitting of the data using Quantum-enhanced Support Vector Machine (QSVM) is very efficient and all the algorithms used which in turn helps in producing better results.

Table.1. Colon Cancer Average Accuracy

Algorithm	Average Accuracy (%)
PCA	81
GA	85
HFCM	89

The Table.1 displays the results combined with Quantum-enhanced Support Vector Machine (QSVM) and produces the results of algorithm. This dataset is similar to the colon gene dataset for evaluation of data. It contains expression levels of 2000 gene samples taken in sixty-two different verities samples. For each sample it is point out, whether it is taken from a tumor biopsy or not. The average accuracy displayed by the Principal Component Analysis (PCA) of is 81% accuracy, Genetic Algorithm (GA) of 85% and Hierarchical Fuzzy C Means (HFCM) is 89%. The finally the results reflects that the maximum of accuracy rate in HFCM is combined with QSVM in Colon Cancer datasets.

The numbers and descriptions for the different levels of normal and abnormal genes are also given. This dataset is used in many different levels research and produces various results. It can be used in two different ways: the total samples can either be treated as records in a high dimensional space or they can treat the accurate genes as records with 62 attributes. The clustering methods of the Average Accuracy rate of Colon dataset is given below,

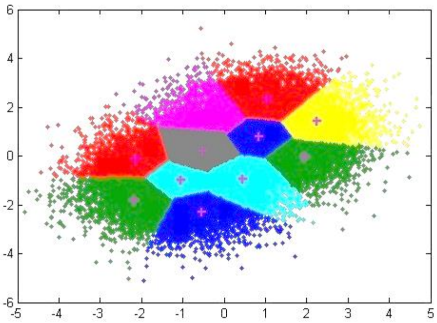

Fig.1. Average Accuracy Rate of Colon Dataset

The Fig.1 shows the average accuracy rate of colon dataset. The lung cancer dataset involves gene expression datasets of 12,533 genes taken from 181 lung tissue samples along with 31 are become the malignant pleural mesothelioma (MPM) and the remaining 150 are of adenocarcinoma (ADCA).

Table. 2. Average Accuracy of Lung Cancer Prediction

Algorithm	Average Accuracy (%)
PCA	81
GA	85
HFCM	91

The Table.2 displays the results combined with Quantum-enhanced Support Vector Machine (QSVM) and produces that average accuracy to display the Principal Component Analysis (PCA) of 81% of the accuracy, Genetic Algorithm (GA) of 85% and Hierarchical Fuzzy C Means (HFCM) of 91%. The final results demonstrate that the maximum of accuracy rate in HFCM combines with the QSVM in Lung Cancer datasets.

This dataset is similar to the lung gene evaluation of dataset and it contains expression levels of 12,533 genes taken in 181 different samples. For each sample it is indicated if it is obtained from a tumor biopsy. The numbers and descriptions for the different levels of genetic values are processed and also results given. This dataset is used in many of the problems and produces different results based on the gene expression data. It can be used in two ways, the 181 samples can be treated as records staggeringly high in space, or the genes can be treated as records with 181 attributes. The Fig.3 shows the average accuracy rate of Lung dataset below:

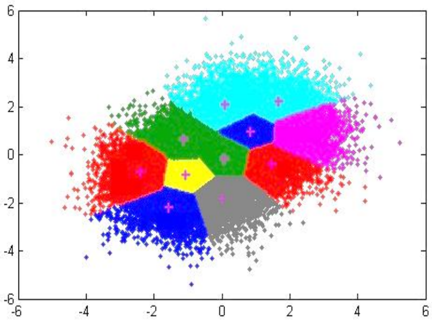

Fig.3. Average Accuracy Rate of Lung Dataset

The Table.3 shows the average accuracy rates of both Colon and Lung datasets. The comparison in Table.3 is done between the Colon and Lung datasets implementation and split of data using QSVM and implements to the PCA, GA and HFCM. The results depict that the proposed of QSVM and HFCM has achieved more accurate results when compared with the other two algorithms using two datasets.

Table.3. Colon and Lung Average Accuracy

Algorithm	Average accuracy (%) of Colon	Average accuracy (%) of Lung
PCA	81	81
GA	85	85
HFCM	89	91

5. CONCLUSION

In this research paper there are three data mining algorithms used to compare the cancer data for finding the accuracy. This study identifies that QSVM and HFCM has achieved more accurate results. The comparison of QSVM and HFCM loss of information is lower than in PCA algorithm. The comparison of QSVM and HFCM with relation to time consuming is higher in GA algorithm and in comparison, to both the algorithm QSVM and HFCM is more efficient to produce accurate results. Better results are also produced in both the colon and lung datasets.

REFERENCES

[1] Y. Wang and F. Makedon, “Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification using Microarray Data”, *Proceedings of IEEE International Conference on Computational Systems Bioinformatics*, pp. 497-498, 2004.

[2] C. Saravanabhavan, T. Saravanan, D.B. Mariappan and K.M. Baalamurugan, “Data Mining Model for Chronic Kidney Risks Prediction Based on using NB-CbH”, *Proceedings of International Conference on Advance Computing and Innovative Technologies in Engineering*, pp. 1023-1026, 2021.

[3] D. Huang, Y. Quan, M. He and B. Zhou, “Comparison of Linear Discriminant Analysis methods for the Classification of Cancer based on Gene Expression Data”, *Journal of Experimental and Clinical Cancer Research*, Vol. 28, No. 1, pp.1-8, 2009.

[4] Y. Han and L. Yu, “Margin based Sample Weighting for Stable Feature Selection”, *Proceedings of International Conference on Web-Age Information Management*, pp. 680-691, 2010.

[5] V. Chang, B. Gobinathan, A. Pinagapani and S. Kannan, “Automatic Detection of Cyberbullying using Multi-Feature based Artificial Intelligence with Deep Decision Tree Classification”, *Computers and Electrical Engineering*, Vol. 92, pp. 1-21, 2021.

[6] P. Valarmathie, T. Ravichandran and Dinakaran, K, “Survey on Clustering Algorithms for Microarray Gene Expression Data”, *European Journal of Scientific Research*, Vol. 69, No. 1, pp. 5-20, 2012.

[7] S. Hengpraprohm and P. Chongstitvatana, “Feature Selection by Weighted-SNR for Cancer Microarray Data Classification”, *International Journal of Innovative Computing, Information and Control*, Vol. 5, No. 12, pp. 4627-4636, 2009.

[8] N. Singh, K. Guliani and P. Prabhat, “Comparison of Different Filtering Approaches on Gene Expression Data for Clustering”, *International Journal of Engineering Research and Technology*, Vol. 2, No. 5, pp. 815-818, 2013.

[9] R. Anand, S. Ravichandran and S. Chatterjee, “A New Method of Finding Groups of Coexpressed Genes and Conditions of Coexpression”, *BMC Bioinformatics*, Vol. 17, No. 1, pp. 1-14, 2016.

[10] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas and E. Adebiyi, “Clustering Algorithms: their Application to Gene Expression Data”, *Bioinformatics and Biology Insights*, Vol. 10, pp. 237-253, 2016.

[11] H.W. Nies, K.M. Daud, M.A. Remli, M.S. Mohamad, S. Deris, S. Omatu, S. Kasim and G. Sulong, “Classification of Colorectal Cancer using Clustering and Feature Selection Approaches”, *Proceedings of International Conference on Practical Applications of Computational Biology and Bioinformatics*, pp. 58-65. 2017.

[12] M.A. Remli, K.M. Daud, H.W. Nies, M.S. Mohamad, S. Deris, S. Omatu, S. Kasim and G. Sulong, “K-Means Clustering with Infinite Feature Selection for Classification Tasks in Gene Expression Data”, *Proceedings of International Conference on Practical Applications of Computational Biology and Bioinformatics*, pp. 50-57, 2017.

[13] X. Yu, G. Yu and J. Wang, “Clustering Cancer Gene Expression Data by Projective Clustering Ensemble”, *PloS One*, Vol. 12, No. 2, pp. 1-21, 2017.

[14] S.M. Ayyad, A.I. Saleh and L.M. Labib, “Gene Expression Cancer Classification using Modified K-Nearest Neighbors Technique”, *Biosystems*, Vol. 8, No. 2, pp. 41-51, 2019.

[15] N. Naicker, T. Adeliyi and J. Wing, “Linear Support Vector Machines for Prediction of Student Performance in School-Based Education”, *Mathematical Problems in Engineering*, Vol. 2020, pp. 1-7, 2020.