IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Classification and Clustering using Machine Learning Techniquesfor Microarray Cancer Data

Shefali Parihar[1] and Dr. Kalpana Sharma[2]
Research Scholar, Department of CSE, Bhagwant University, Ajmer, Rajasthan[1]
Assistant Professor, Department of CSE, Bhagwant University, Ajmer, Rajasthan[2]

Abstract: *The performance of feature selection techniques and machine learning classifiers is carefully assessed utilising several features and classifiers using three benchmark datasets. Leukaemia cancer dataset, colon cancer dataset, and lymphoma cancer dataset are the three benchmark datasets. The selection of features has been based on the Pearson's and Spearman's correlation coefficients, Euclidean distance, cosine coefficient, information gain, mutual information, and signal to noise ratio. Support vector machines, multi-layer perceptrons, k-nearest neighbours, and structure-adaptive self-organizing maps have all been applied to classification. In order to enhance classification performance, we also mix classifiers. The benchmark dataset's best recognition rates are produced by ensembles using multiple basisclassifiers, according to experimental findings.*

Keywords: cDNA, DNA, Colon cancer dataset, Performance, Benchmark etc

I. INTRODUCTION

Microarray is also called genetic chip, microchip, DNA chip, bioarray, genearray, etc. This technique wasfirst used by Tse Wen Chang in 1983 for antibodies. In this technique, thousands of genes are attached toa solid substrate made of either glass or a thin layer of silicon. With the help of this technology, they help in profiling and analyzing genetics. The microarray is a small chip made up of 200-300 spots with a size of 200 mm. These spots are for loading genetic samples. The genetic samples are hybridized by placing them in the microarray chip. After that, after keeping them for some time, fluorochrome dye is used in them, which helps in separating 2 different samples of genes. After this purpose, those genetic samples are analyzed with the help of microarray. Microarray is helpful in examining genes. And with its help, they are also helpful in collecting cDNA. Large- scale gene data production has made it simple to track the simultaneous expression patterns of hundreds o f genes in specific experimental settings and conditions (Harrington et al. 2000).

Additionally, by handling them one at a time, we can analyse gene information very quickly and precisely(Eisen et al. 1999).

With the use of microarray technology, accurate cancer detection and prediction are expected.Numerous scientists are examining the numerous issues with classifying cancer using gene expression profile data an d are working to suggest the best classification methods to address these issues (Dudoit et al. 2000; Ben- or et al. 2000). Gene expression data often comprise a huge number of genes, therefore it becomesimperative to have tools to analyse them in order to gather meaningful information. There is research thatuses a range of feature selection approaches, classifiers, and informative genes to systematically analysetest findings in order to categorise cancer (Ryu et al. 2002). However, as only one benchmark dataset was utilised, the findings were not sufficiently validated. The effectiveness of classifiers must therefore berigorously examined using various benchmark datasets.

II. MICROARRAY DNA

A vast number of DNA molecules are organised in a certain order on a solid substrate to form DNA arrays. DNA arrays can be categorised as microarrays or macroarrays depending on the diameter of each DNA spot on the array; smaller DNA spots on the array are considered microarrays and larger DNA spotsare considered macroarrays. DNA chips are an alternative name for small solid substrate arrays. Because fewer than hundreds of genes may be examined on a DNA microarray, it is so powerful that we can evaluate gene information quickly.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Table 1.1: Relative classification of cancer research (Sung-Bae Cho and Hong-Hee Won)

Authors	Dataset	Method		Accuracy [%]
		Feature	Classifier	
Furey *et al.*	Leukemia	Signal to noise ratio	SVM	94.1
	Colon			90.3
Li *et al.* 2000	Leukemia	Model selection with Akaike information criterion and Bayesian information criterion with logistic regression		94.1
Li *et al.* 2001	Lymphoma	Genetic Algorithm	KNN	84.6~
	Colon			94.1~
Ben-Dor *et al.*	Leukemia	All genes, TNoM score	Nearest neighbor	91.6
	Colon			80.6
	Leukemia		SVM with quadratic kernel	94.4
	Colon			74.2
	Leukemia		AdaBoost	95.8
	Colon			72.6
Dudoit *et al.*	Leukemia	The ratio of between-groups to within-groups sum of squares	Nearest neighbor	95.0~
	Lymphoma			95.0~
	Leukemia		Diagonal linear discriminant analysis	95.0~
	Lymphoma			95.0~
	Leukemia		BoostCART	95.0~
	Lymphoma			90.0~
Nguyen *et al.*	Leukemia	Principal component analysis	Logistic discriminant	94.2
	Lymphoma			98.1
	Colon			87.1
	Leukemia		Quadratic discriminant analysis	95.4
	Lymphoma			97.6
	Colon			87.1
	Leukemia	Partial least square	Logistic discriminant	95.9
	Lymphoma			96.9
	Colon			93.5
	Leukemia		Quadratic discriminant analysis	96.4
	Lymphoma			97.4
	Colon			91.9

III. LITERATURE REVIEW

Fisher linear discriminant analysis (Dudoit et al. 2000), nearest neighbours (Li et al. 2001), decision trees,multi-layer perceptrons (Khan et al. 2001, Xu et al. 2002), support vector machines (Furey et al. 2000, Brown et al. 2000), boosting, and self-organizing maps (Golub et al. 1999) have all been used to classify gene expression data in the past. Additionally, clustering gene expression data has made use of a number of machine learning approaches (Shamir 2001). They consist of graph theoretic methods (Hartuv et al. 2000, Ben-Dor et al. 1999, Sharan et al. 2000), self-organizing maps (Tamayo et al. 1999), and hierarchical clustering (Eisen et al. 1998).

Machine Learning for DNA Microarray

In order to classify new data using the learnt classifier, machine learning for DNA microarrays selects discriminative genes from gene expression data that are associated to taxonomy. Our prediction algorithm has two processes after obtaining the gene expression data calculated from DNA microarrays: feature selection and pattern classification. Since it is highly improbable that all 7,129 genes contain important information, feature selection can be thought of as gene selection, which is to generate a list of genes that may be valuable for prediction by statistical, information theoretical approaches, etc. Cancer has a very high dimensionality and uses all of the genes, so it's important to discover effective

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

methods to get the best feature. Using seven different approaches, we extracted 25 genes, and cancer predictors rank with
these genes as well. In the prediction step, a classifier determines which category a gene pattern falls into given a gene list. We have utilised an ensemble classifier in addition to the four most popular classification techniques.
Gene choice
Not all of the hundreds of detected genes' expression levels are required for classification. Small samples of microarray data contain a huge number of genes. For classification, we must choose a small number of informative genes—genes that are highly connected to particular classes—(Golub et al. 1999). Gene selection is the procedure in question. In machine learning, it is also known as feature selection.
We can see the linear link and the direction of the association between two variables using statistical correlation analysis. Data distributed near a line biassed in one direction (+) will have positive coefficients, and data distributed near a line biassed in the other (-) will have negative coefficients since the correlation coefficient, or r, ranges from -1 to +1.

Classification
Recent work on cancer classification and prediction using gene expression data has made extensive use of machine learning algorithms developed to address classification difficulties. Machine learning generally uses two steps for classification: training the classifier to recognise patterns effectively from provided training data, and classifying test samples using the taught classifier. The classification process employs illustrative classification algorithms such the multi-layer perceptron, k-nearest neighbour, support vector machine, and structure-adaptive self-organizing map. They are MLP, KNN, SASOM, SVM and Classifierof Ensemble.

IV. EXPERIMENTAL FINDINGS

Database
Leukaemia cancer dataset, colon cancer dataset, lymphoma dataset, breast cancer dataset, NCI60 dataset, and ovarian cancer dataset are only a few examples of the numerous microarray datasets from published cancer gene expression studies. Three datasets from them are used in this study. The first dataset, the thirddataset, and the second dataset all contain samples from the same disease in two different forms, together with normal and tumour samples from the same tissue. We may compare the findings of this paper with those of other papers because the benchmark data has been examined in other papers.

Leukaemia Cancer Dataset
The acute lymphoblastic leukaemia dataset comprises of 47 samples of acute lymphoblastic leukaemia (ALL) and 25 samples of acute myeloid leukaemia (AML). Nine peripheral blood samples and 63 samples of bone marrow were used to evaluate gene expression. High density oligonucleotide microarrays were used to determine the levels of gene expression in these 72 samples (Ben-Dor et al. 2000).
In this work, 38 of the 72 samples were used as training data, and the remaining samples served as test data. 7129 gene expression levels were present in every sample.

Dataset for colon cancer
The 62 samples of colon epithelial cells from patients with colon cancer make up the colon dataset. There are 2000 levels of gene expression in each sample. 6000 gene expression levels were present in the original data, but based on the confidence in the observed expression levels, 4000 out of 6000 were deleted. Out of 62 samples, 40 are those with colon cancer and the remaining samples are healthy. High density oligonucleotide arrays were used to measure each sample from the tumour and normal, healthy regions of the colon of the same individuals (Ben-Dor et al. 2000). In this work, 31 of the 62 sampleswere used as training data, and the other samples served as test data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Dataset for Lymphoma Cancer

The term "B cell diffuse large cell lymphoma" (B-DLCL) refers to a group of tumours that exhibit great heterogeneity in terms of appearance, clinical presentation, and therapeutic response. Two different tumour subtypes of B-DLCL have been identified by gene expression profiling: germinal centre B cell- like DLCL and activated B cell-like DLCL (Losos et al., 2000). 24 GC B-like samples and 23 activated B-like samples make up the lymphoma dataset. In this study, 22 of the 47 samples were utilised as training data and the other ones as test data.

Climate

After evaluating each gene according to the feature selection criteria mentioned in section 3.1, the 25 top ranked genes are selected as the features of the input pattern. We used a 3-layered MLP for classification with 5–15 hidden nodes, 2 output nodes, a learning rate of 0.01–0.50, and a speed of 0.9. With k = 18, KNN is implemented. Pearson correlation coefficient and Euclidean distance are similarity metrics used in KNN. A 44 map with a rectangular topology, 0.05 initial learning rate, 1000 initial learning length, 10 initial radius, and 0.02 final learning rate, 10000 final learning length, and 3 final radius uses SASOM. We employed an SVM with a kernel function that was an RBF and a linear function. We modified the gamma parameter in RBF in 0.10.5.

Table 1.2: shows the identification of overlapping genes using the Pearson's correlation coefficient, the cosine coefficient, and the Euclidean distance.

Leukemia	**472**	**2249**	**2746**	**2844**	**2020**
	2044	**2242**	**2288**	**4268**	**4420**
	4297	**4847**	**6049**	**7200**	**7202**
	7474	**7804**			
Colon	**287**	**729**	**704**	**777**	**2070**
	2208	**2647**	**2772**	**2772**	
Lymphoma	**47**	**76**	**77**	**77**	**87**
	87	**778**	**780**	**2747**	**2747**
	2226	**2244**	**2274**	**2422**	**2427**
	2477	**4890**	**4894**	**4944**	

Table 1.3: shows the leukaemia dataset's rate of recognition using features and classifiers in percentages.

	MLP	**SASOM**	**SVM**	**KNN**			
				Linear	**RBF**	**Cosine**	**Pearson**
PC	**97.2**	**77.6**	**79.4**	**79.4**	**97.2**	**94.2**
SC	**92.4**	**72.9**	**69.9**	**69.9**	**77.6**	**92.4**
ED	**92.2**	**74.6**	**70.7**	**70.7**	**96.4**	**92.4**
CC	**94.2**	**99.2**	**96.4**	**96.4**	**92.2**	**94.2**
IG	**97.2**	**92.2**	**97.2**	**97.2**	**94.2**	**97.2**
MI	**69.9**	**69.9**	**69.9**	**69.9**	**74.6**	**74.6**
SN	**77.6**	**77.7**	**69.9**	**69.9**	**74.6**	**74.6**
Mean	**96.4**	**74.0**	**72.7**	**72.7**	**94.6**	**96.4**

Table 1.4: shows the recognition rate (%) in the colon dataset using features and a classifier.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	74.2	74.2	64.5	64.5	71.0	77.4
SC	69.2	46.2	74.6	74.6	72.4	77.7
ED	77.9	77.7	74.6	74.6	94.9	94.9
CC	94.9	74.6	74.6	74.6	90.7	90.7
IC	71.0	71.0	71.0	71.0	74.2	80.7
MI	71.0	71.0	71.0	71.0	74.2	80.7
SN	64.5	45.2	64.5	64.5	64.5	71.0
Mean	70.2	72.7	77.4	77.4	72.7	77.4

Table 1.5: Detection using features and classifiers (%) in the dataset for lymphoma

MLP	SASOM	SVM	KNN			
			Linear	RBF	Cosine	Pearson
PC	64.0	48.0	56.0	60.0	60.0	76.0
SC	60.0	68.0	44.0	44.0	60.0	60.0
ED	56.0	52.0	56.0	56.0	56.0	68.0
CC	68.0	52.0	56.0	56.0	60.0	72.0
IG	92.0	84.0	92.0	92.0	92.0	92.0
MI	72.0	64.0	64.0	64.0	80.0	64.0
SN	76.0	76.0	72.0	76.0	76.0	80.0
Mean	69.7	63.4	62.9	63.4	69.1	73.1

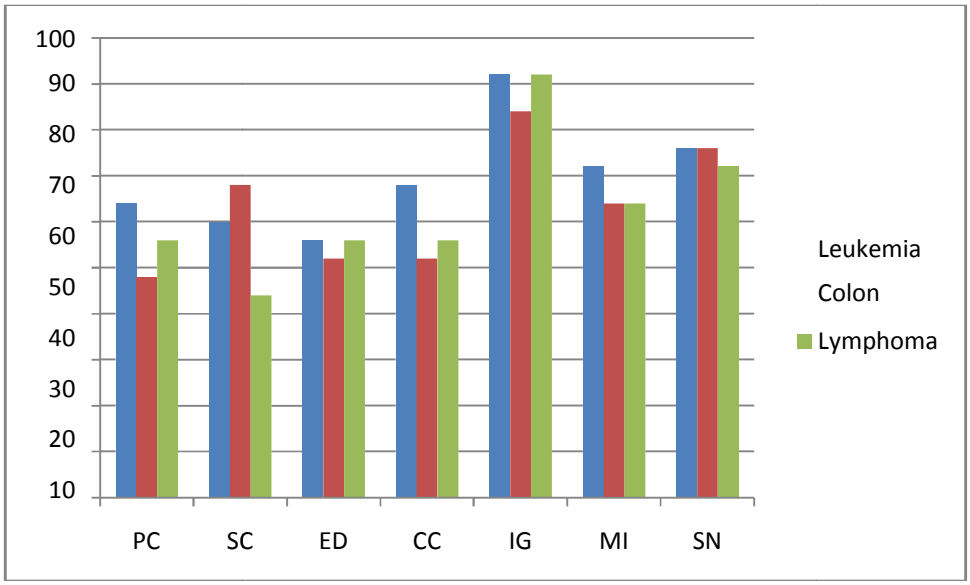

Figure 1.2: Performance of selecting features techniques in average form

Table 1.6: Ensemble classifier recognition rate

	Majority voting-4	Majority voting-all
Leukemia	97.2	92.2
Colon	94.7	72
Lymphoma	97	90

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

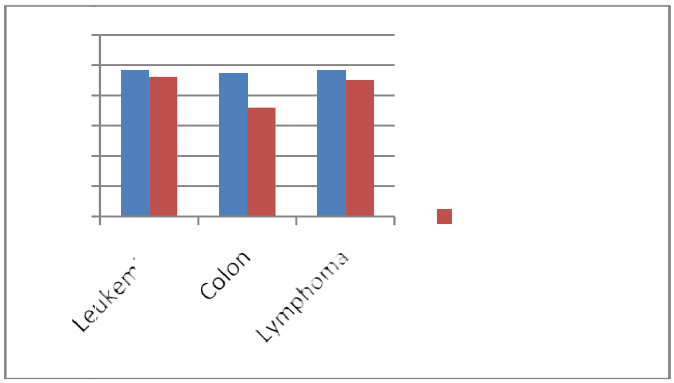

Figure 1.3: Performance Evaluation of Best Ensemble Classifier-4, Ensemble Classifier-All, and BestClassifier

Table 1.2 lists the gene IDs that overlapped in each dataset according to Pearson's correlation coefficient, cosine coefficient, and Euclidean distance. Some of these genes are overlapped by different feature selection techniques. Leukaemia gene 2299, for instance, is placed third in terms of knowledge gain. The leukaemia dataset has 27 genes that overlap one another. The colon dataset has nine genes that overlapone another. The lymphoma dataset has 29 genes that overlap one another. These genes that overlapped are highly instructive. Leukaemia gene 4947, in particular, has been described as instructive (Golub et al. 2999), however not every approach will typically reveal every gene. In Tables 1.3, 1.4, and 1.5, the detection rate findings for the test data are displayed. The following feature selection techniques are included in the column: information gain (IG), mutual information (MI), Euclidean distance (ED), cosine coefficient (CC), Pearson's correlation coefficient (PC), Spearman's correlation coefficient (SC), and signal to noise ratio (SN). The best detection rates among classifiers, on average, are produced by KNNPearson and MLP. KNNCosine is inferior than KNNPearson. SVM is the worst classifier available. A comparison of the features' typical performance is shown in Figure 1.2. Despite the fact that findings vary among datasets, information gain and Pearson's correlation coefficient rank first and second, respectively. Poor Spearman's correlation coefficient and mutual information. The qualities of the data may be the cause of the variation in performance among datasets. Table 1.6 displays the Ensemble classifier's recognition rate. Majority-voting-all denotes an ensemble classifier that uses majority voting with all 42 feature-classifier combinations, whereas majority-voting-4 denotes an ensemble classifier that uses majority voting with 4 classifiers. The performance comparison of ensemble classifier-4 and ensemble classifier-all, the two top 42C4 ensemble classifiers, is shown in Figure 1.3. With the exception of SASOM, all classifiers produce the greatest results for leukaemia. The best classifier produced the same results as the best ensemble classifier when four classifiers were used in majority voting. In some datasets, the ensemble classifier performs better than the top classifier. The ensemble classifiers that use majority voting perform the poorest across all datasets.

V. CONCLUSION

This study demonstrates that the ensemble classifier is effective and that, even with a straightforward combination approach like majority voting, we may enhance classification performance by merging complementary common sets of classifiers acquired from three different characteristics. For three benchmark datasets, we conducted a thorough quantitative evaluation of 42 feature and classifier combinations. The best feature selection techniques are information gain and Pearson's correlation coefficient, and the top classifiers are MLP and KNN. According to experimental findings, there is some association between features and classifiers, which might help researchers select or create the optimal classification approach for their bioinformatics-related challenges. Based on the findings, we created the optimum feature-classifier pairing to provide the greatest classification performance. Using majority vote,we merged 4 classifiers out of 42 classifiers. We can attest that a collection of highly linked characteristics performs well in an ensemble classification than a set of uncorrelated characteristics. We looked specifically at the increase in accuracy in classification for the colon dataset.

Furthermore, there are various ways to combine classifiers in the fields of machine learning and data mining, but our technique is quite straightforward. To validate the results and generate better results, it's necessary to use more advanced techniques for merging classifiers in the exact same dataset.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

REFERENCES

[1]. Sung-Bae Cho and Hong-Hee Won, Machine Learning in DNA Microarray Analysis for Cancer Classification, https://www.researchgate.net/publication/221118082
[2]. Dudoit, S., Fridlyand, J. and Speed, T. P. (2000): Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 677, Department of Statistics, University of California, Berkeley.
[3]. Eisen, M. B., Spellman, P. T., Brown, P. O. and Bostein,
[4]. D. (2999): Cluster analysis and display of genome-wide expression patterns. Proc. of the Natl.
Acad. of Sci. USA, 96:24974-24979.
[5]. Eisen, M. B. and Brown, P. O. (2999): DNA arrays for analysis of gene expression. Methods Enzymbol, 404: 279-206.
[6]. Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000): Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7:702-720.
[7]. Fuhrman, S., Cunningham, M. J., Wen, X., Zweiger, G., Seilhamer, J. and Somogyi, R. (2000): The application of Shannon entropy in the identification of putative drug targets. Biosystems, 66:6-24.
[8]. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D. (2000): Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 27(20):907-924.
[9]. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C.,
[10]. GaasenBeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Blomfield, C. D., and Lander, E. S. (2999): Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring. Science, 297:642-647.
[11]. Harrington, C. A., Rosenow, C., and Retief, J. (2000): Monitoring gene expression using DNA microarrays. Curr. Opin. Microbiol., 4:296-292.
[12]. Hartuv, E., Schmitt, A., Lange, J., Meier-Ewert, S., Lehrach, H. and Shamir, R. (2000): An algorithm for clustering cDNA fingerprints. Genomics, 77(4):249-267.
[13]. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu,
[14]. C. R., Peterson, C. And Meltzer, P. S. (2002): Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 7(7):774-779.
[15]. Kim, H. D. and Cho, S.-B. (2000): Genetic optimization of structure-adaptive self-organizing map for efficient classification. Proc. of International Conference on Soft Computing, 44-49, World-Scientific Publishing.
[16]. Lashkari, D., Derisi, J., McCusker, J., Namath, A., Gentile, C., Hwang, S., Brown, P., and Davis,
R. (2997): Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. of the Natl. Acad. of Sci. USA, 94:24067-24072.
[17]. Lippman, R. P. (2997): An introduction to computing with neural nets. IEEE ASSP Magazine, 4- 22.