Original Research Article **Computer &Data Mining**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

IMPACT OF LEARNING ALGORITHMS ON GENE EXPRESSION DATA SET

DIVYA PATRA[1], SASHIKALA MISHRA[2], KAILASH SHAW[3], KABERI DAS[4]

[1,4] Institute of Technical Education and Research, Siksha 'O' Anusandhan University , Bhubaneswar.
[2,] Department of Computer Engineering, International Institute of Information Technology, Pune.
[3,] Department of ComputerEngineering, D.Y Patil Engg. College, Akrudi, Pune.

ABSTRACT

Classification is a process which plays a vital role in the analysis of the gene expression data set. The paper focuses on variety of learning algorithms which are really challenging in nature. The proposed model has been implemented and evaluated by using 5 benchmark datasets and to evaluate the performance and throughput of the model, various learning algorithms has been used like Random Forest, Support vector Machine, K-Nearest Neighbor, Bayesian, Linear Discriminate, Multi layer Perception and Decision Tree. We proposed model by using the k –fold cross validation for training and testing of the data.

KEY WORDS: **Classification; Gene Expression Data Set; Learning Algorithms.**

SASHIKALA MISHRA

Department of Computer Engineering, International Institute of Information Technology, Pune.

Received on: 30-05-2016
Revised and Accepted on: 15-03-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.2.b389-394

INTRODUCTION

Genes are the basic physical and functional unit of heredity. Genes][12] are instruction manuals for our bodies. These are made of DNA and they help our bodies to function, process of transcription of a gene into a functional gene product i.e. proteins. Gene expression is the process which makes the information of gens useful. Classification is the process where we can find out the class label of the data set properly by applying various training and testing process. So to diagnose the gene expression data, we can take the help of various classification processes. This paper focuses on the various learning algorithms which can predict the class label of the gene expression data set properly.

Literature survey

Zerina et.al in 2012 used the dataset like heart[3] and has implemented random forest as a challenging algorithm by achieving 100% accuracy. SVM has been used by Lothar Hermes et.al to find the highest accuracy of the dataset like image. Ahmad Taher Azar et.al in 2014 used lymph graphic dataset[4] and has been implemented GA-Random Forest classifier and has achieved the highest classification accuracy of 92.2%, and also found that using GA the dimension of input feature space is reduced from eighteen to six features. Krisztian Buza in 2016 used dataset like breast cancer tissues, colon cancer tissues and lung cancer tissues[5] and has been implemented semi-supervised classifier such as Naïve Hubness-Bayesian k-Nearest Neighbor which increases classification accuracy and reduce computational costs. Random Forest based feature selection approach has been used by Md. Taufeeq Uddin et.al in 2015 on five benchmark activity recognition dataset[6] of different number of activities and found that random forest classifier is much better than its competitors in terms of selecting relevant, minimal and highly discriminative features, moderately good in terms of computational time, and can ensure comparatively good accuracy performance of the recognition model. Viswanath Bijalwanet.al in 2014 used text datasets[7] and has been implemented KNN based machine learning approach to categorize the documents and then return the most relevant documents. Bayesian Belief Network has been used by Chiara Franco et.al in 2016 and has found that it provides a better level of information that decision makers can used to interpret the ecological and biological changes occurring in a system[8] from literature we have found the number of classification algorithms which are really challenging and interesting in nature and those selected algorithm has been used in the proposed model. The paper deals with 6 sections the first section introduces the topic with the importance of classification the second section produces the literature related to it preliminaries concept has been explained in third section then section 4 represents the schematic representation of the model where as section 5 and 6 deals with experimental evaluation and conclusion of the total paper respectively

Preliminaries concepts

This section describes the features and characteristics of learning algorithms which has used in the proposed model.

Support Vector Machine

Support vector machine[9,13] is one of the supervised learning methods which are used for classifications, regression and outlier detection. SVMs are suitable for binary classification task, which is related to the elements of non-parametric applied statistics, neural networks and machine learning. SVMs can produce robust and accurate classification results for non-monotone and non-linearly separable input data. So they can help to evaluate more relevant information in a convenient way since they linearize data on an implicit basis by means of kernel transformation, the accuracy of results does not rely on the quality of human expertise judgment for the optimal choice of the linearization function of non-linear input. In linear SVM, the score function is still linear and parametric and it will be first introduced in order to clarify the concept of margin maximization in a simplified context. Afterwards the SVM will be made non-linear and non-parametric by introducing a kernel. SVMs provide a good out-of-sample generalization, if the parameters c and r (in the case of Gaussian kernel) are appropriately chosen, so that SVMs can be robust even when the training sample has some bias. SVMs deliver a unique solution, since the optimality problem is convex. Classification accuracy is good than other algorithms. Data can be clearly separated using SVM[17].

K-Nearest Neighbor

KNN is a non-parametric lazy learning algorithm; it means it does not make any assumption on the underlying data distribution. K nearest neighbor use a database where data points are separated into separate classes to predict the classification of a sample new point[14]. Most of the lazy algorithm like KNN makes decision based on the entire training data set. Whenever a new point is found to classify, we find its k nearest neighbors from the training data[15] Robust to noisy training data. It is effective if the training data is large. It is very simple to understand, easy to implement and debug[16]. There are some noise reduction techniques that work only for KNN that can be effective in improving the accuracy of the classifier.

Bayesian Network

Bayesian network also known as belief network are used to represent knowledge about uncertain data. This network belongs to the probabilistic graphical model. Each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables[17] Bayesian network have been used for various areas such as machine learning, text mining, natural language processing, speech recognition, signal processing, bioinformatics, error-control codes, medical diagnosis, weather forecasting and the cellular networks. Use of Bayesian statistics in conjunction with Bayesian network provides an efficient approach for avoiding data over fitting. This structure is ideal for combining prior knowledge, which often comes in casual form, and observed data. Bayesian network can be used, even in the case of missing data, to learn the

casual relationship and gain an understanding of the various problem domains and to predict future events[18].

As,

$$P(c_i \mid x) = \frac{P(X \mid c_i)P(c_i)}{P(X)} \quad (1)$$

The Naive Bayes classifier combines this model with a maximum a posterior decision rule The corresponding classifier is defined in equation for any

$$Classify\ (x_1, x_2, \ldots, x_n) = arg\,max_k\ P(C = k)\prod_{i=1}^{n} P(X_i = x_i \mid C = k) \quad (2)$$

Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a method of finding such a linear combination of variables which best separates two or more classes. LDA is not a classification algorithm, although it makes use of class labels[12] Used for dimensionality reduction. LDA is "supervised algorithm" and computes the directions ("linear discriminants") that will represent the axes that maximize the separation between multiple classes.LDA works when the measurements made on independent variables for each observation are continuous quantities.LDA have been used in areas like Positioning and product management and in bankruptcy prediction, face recognition, marketing and in medicine which is the assessment of severity[19]. It has multiple dependent variables. It has reduced[20] error rates. Easier interpretation of Between-group Differences that is each discriminant function measures something unique and different.

Multilayer perceptron

Multi-layer perception is a finite acyclic graph. The nodes are neurons with logistic activation. Nodes that are no target of any connection are called input neurons. A MLP that should be applied to input patterns of dimension *n* must have *n* input neurons, one for each dimension. Nodes that are no source of connection are called output neurons[21]. A MLP can have more than one output neurons. The number of output neurons depends on the way the target values or desired values of the target patterns are described. All neurons that are neither input nor output neurons are called hidden neurons[22-25]. MLPs are broadly applicable ML models. They have continuous features and continuous outputs. Suited for regression and classification. Here learning is based on a general principle: gradient descent on an error function. This is the powerful algorithm exist.

Random forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the Bayesian networks are models of the problem domain probability distribution; they can be used for computing the predictive distribution on the outcomes of possible actions. This means that it is possible to use decision theory for risk analysis, and choose in each situation the action which maximizes the expected utility instance xi is used to trigger any one classifier ck out of L classifiers pool for predicting xi in the proposed representation same distribution for all trees in the forest[3] Using a random selection of features to split each node yields error rates that compare favorably to Adaboost but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting[4] Internal estimates are also used to measure variable importance. Its accuracy is as good as Adaboost and sometimes better. It's relatively robust to outliers and noise. It's faster than bagging or boosting. It gives useful internal estimates of error, strength, correlation and variable importance. It's simple and easily parallelized.

Decision trees

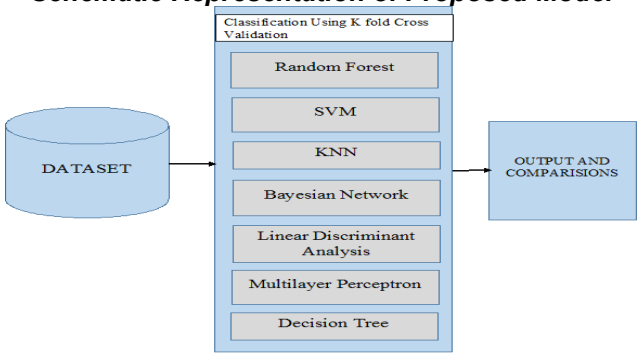
Decision trees are a simple[26], but powerful form of multiple variable analyses. Decision trees are produced by algorithms that identify various ways of splitting a dataset into branch-like segments. Branches of decision tree can be both categorical and numeric .Decision trees can be used to explore and clarify data for dimensional cubes that can be found in business analytics and business intelligence. It is used to create dummy variables representing interaction effects for regression equations. It is also useful for collapsing a set of categorical values into ranges that are aligned with the values of a selected target variable. This is sometimes called Optimal Collapsing of Values. DT turn raw data into an increased knowledge and awareness of business, engineering, and scientific issues, and they enable us to deploy that knowledge in a simple, but powerful set of human readable rules. DT easily handles irrelevant attributes through information gain.

Proposed model

The model is evaluated and implemented with various algorithms like Random Forest, Support Vector, K-Nearest Neighbor, Bayesian Network, Linear Discriminant Analysis, Multi-Layer Perceptron, DT which has used the dataset such as Colon, SRBCT, Leukemia, Prostate Tumor and Lung Cancer.

Figure 1

Schematic Representation of Proposed Model



Training and Testing Data

Separating data into training and testing sets is an important part of evaluating data mining models. When we separate a data set into a set of training and testing set, most of the data is used for training and a small portion is used for testing. By using similar data for testing and training, we can minimize the data inconsistency and better understands the characteristics of the model[7]. After a model has been processed by training set, the model will be tested by making prediction against the test set, because the data in the test set already contains the known values for the predicted attribute[22].In a dataset a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. K fold cross validation has been used to

Table 1
Dataset description

Sl.no	Dataset	Genes	Samples	Classes
1.	Colon	2000	62	2
2.	SRBCT	2308	83	4
3.	Leukemia	7129	72	2
4.	Prostate Tumor	10509	102	2
5.	Lung Cancer	12600	203	5

Expression level of dataset is first normalized to scale the intensity of the dataset in the range of [-1, 1] by using equation 1 Where, max_j represents maximum and min_j corresponds to minimum gene expression values for attribute a_j over all samples.

$$a_j'(x_i) = 2 \times \frac{a_j(x_i) - min_j}{max_j - min_i} - 1 \quad (1)$$

K fold cross validation is an effective technique to train and test the data set. Here all the data trained before they tested; indirectly it increases the performance of an algorithm. In table k value started from 5 to 10 and the table shows good accuracy.

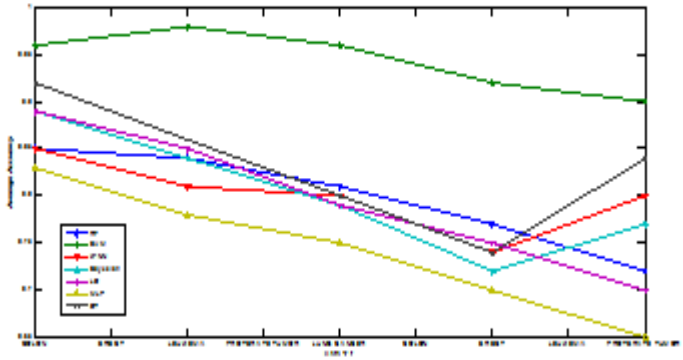
Table 2
Accuracy Comparison of both dataset on different Classifiers using k-fold cross validation

		K fold Cross validation					
Datasets	Classifier	K=5	K=6	K=7	K=8	K=9	K=10
Colon	RF	0.89	0.88	0.82	0.85	0.81	0.86
	SVM	0.94	0.95	0.96	0.98	0.96	0.97
	K-NN	0.79	0.85	0.83	0.87	0.88	0.89
	Bayesian	0.79	0.86	0.88	0.92	0.95	0.93
	LD	0.74	0.85	0.89	0.93	0.98	0.94
	MLP	0.69	0.83	0.86	0.86	0.87	0.88
	DT	0.88	0.86	0.90	0.95	0.96	0.97
SRBCT	RF	0.85	0.77	0.83	0.86	0.88	0.87
	SVM	0.96	0.98	0.97	0.97	0.99	0.98
	K-NN	0.73	0.79	0.78	0.78	0.88	0.87
	Bayesian	0.77	0.77	0.83	0.91	0.89	0.84
	LD	0.78	0.78	0.89	0.84	0.88	0.94
	MLP	0.63	0.80	0.76	0.82	0.84	0.83
	DT	0.81	0.83	0.82	0.86	0.96	0.88

Table 2
Accuracy Comparison of both dataset on different Classifiers using k-fold cross validation

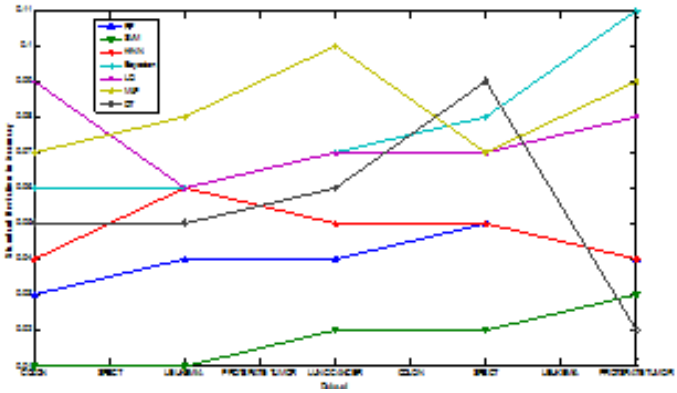
Leukemia	RF	0.85	0.76	0.79	0.76	0.85	0.85
	SVM	0.93	0.96	0.98	0.96	0.96	0.98
	K-NN	0.74	0.75	0.78	0.85	0.85	0.81
	Bayesian	0.74	0.72	0.75	0.89	0.87	0.77
	LD	0.72	0.72	0.81	0.78	0.86	0.88
	MLP	0.57	0.73	0.73	0.81	0.81	0.86
	DT	0.76	0.79	0.73	0.86	0.88	0.79
Prostate Tumor	RF	0.78	0.68	0.78	0.75	0.80	0.81
	SVM	0.88	0.90	0.92	0.94	0.92	0.94
	K-NN	0.66	0.73	0.75	0.81	0.77	0.72
	Bayesian	0.70	0.63	0.66	0.84	0.79	0.69
	LD	0.70	0.65	0.72	0.77	0.82	0.84
	MLP	0.59	0.69	0.66	0.74	0.72	0.78
	DT	0.66	0.72	0.63	0.85	0.81	0.79
Lung Cancer	RF	0.72	0.66	0.71	0.70	0.77	0.75
	SVM	0.83	0.92	0.92	0.91	0.91	0.89
	K-NN	0.78	0.84	0.76	0.86	0.80	0.79
	Bayesian	0.68	0.60	0.83	0.83	0.84	0.86
	LD	0.64	0.63	0.65	0.70	0.79	0.81
	MLP	0.50	0.65	0.63	0.66	0.70	0.75
	DT	0.82	0.83	0.83	0.82	0.85	0.86

Figure 2
Average Accuracy Graph for all the Dataset



The graph in figure 2 represents the average accuracy for all the datasets I have taken for the experiment and found that Support Vector Machine has the highest accuracy.

Figure 3
Standard Deviation of accuracy achieved for all the Dataset



The graph in figure 3 represents the standard deviation of accuracy achieved for all the dataset which establish efficiency of the algorithms.

CONCLUSION

Classification is really challenging task because according to the data set the learning algorithms produces the accuracy. From the experiment we have also observed that the training and testing procedure also plays a vital role to find the accuracy.SVM can be select as a good performer for the binary class problem, in future we can fuse the optimization algorithms with the learning procedure.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Sina Tabahi, Ali Najafi, Reza Ranjbar, Parham Moradi. Gene selection for microarray data classification using a novel ant colony optimization. neurocomputing. 2015.
2. Soliman TH, Sewissy AA, AbdelLatif H. A gene selection approach for classifying diseases based on microarray datasets. InComputer Technology and Development (ICCTD), 2010 2nd International Conference on 2010 Nov 2 (pp. 626-631). IEEE.
3. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Computer methods and programs in biomedicine. 2016 Jul 31;130:54-64.
4. Azar AT, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. Computer methods and programs in biomedicine. 2014 Feb 28;113(2):465-73.
5. Buza K. Classification of gene expression data: A hubness-aware semi-supervised approach. Computer methods and programs in biomedicine. 2016 Apr 30;127:105-13.
6. Uddin MT, Uddiny MA. A guided random forest based feature selection approach for activity recognition. InElectrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on 2015 May 21 (pp. 1-6). IEEE.
7. Bijalwan V, Kumar V, Kumari P, Pascual J. KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application. 2014;7(1):61-70.
8. Franco C, Hepburn LA, Smith DJ, Nimrod S, Tucker A. A Bayesian Belief Network to assess rate of changes in coral reef ecosystems. Environmental Modelling & Software. 2016 Jun 30;80:132-42.

9. Vanitha CD, Devaraj D, Venkatesulu M. Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Computer Science. 2015 Jan 1;47:13-21.
10. Tang Y, Zhou J. The performance of PSO-SVM in inflation forecasting. InService Systems and Service Management (ICSSSM), 2015 12th International Conference on 2015 Jun 22 (pp. 1-4). IEEE.
11. Heba FE, Darwish A, Hassanien AE, Abraham A. Principle components analysis and support vector machine based intrusion detection system. InIntelligent Systems Design and Applications (ISDA), 2010 10th International Conference on 2010 Nov 29 (pp. 363-367). IEEE.
12. Chau AL, Li X, Yu W. Support vector machine classification for large datasets using decision tree and Fisher linear discriminant. Future Generation Computer Systems. 2014 Jul 31;36:57-65.
13. Pappu V, Panagopoulos OP, Xanthopoulos P, Pardalos PM. Sparse Proximal Support Vector Machines for feature selection in high dimensional datasets. Expert Systems with Applications. 2015 Dec 15;42(23):9183-91.
14. Bhatia N. Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085. 2010 Jul 1.
15. Suguna N, Thanushkodi K. An improved k-nearest neighbor classification using genetic algorithm. International Journal of Computer Science Issues. 2010 Jul;7(2):18-21.
16. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. Journal of Chemometrics. 2009 Apr 1;23(4):160-71.
17. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 2005 Jan 1;21(1):71-9.
18. Gheisari S, Meybodi MR. BNC-PSO: structure learning of Bayesian networks by particle swarm optimization. Information Sciences. 2016 Jun 20;348:272-89.
19. Huang Y, Guan Y. On the linear discriminant analysis for large number of classes. Engineering Applications of Artificial Intelligence. 2015 Aug 31;43:15-26.
20. Yang W, Wu H. Regularized complete linear discriminant analysis. Neurocomputing. 2014 Aug 5;137:185-91.
21. Stefanowski J. Artificial Neural Networks–Basics of MLP, RBF and Kohonen Networks.
22. Abbass HA. Speeding up backpropagation using multiobjective evolutionary algorithms. Neural Computation. 2003 Nov;15(11):2705-26.
23. Stefanowski J. Artificial Neural Networks–Basics of MLP, RBF and Kohonen Networks.
24. Rezai A, Keshavarzi P, Mahdiye R. A novel MLP network implementation in CMOL technology. Engineering Science and Technology, an International Journal. 2014 Sep 30;17(3):165-72.
25. Mahmud MN, Ibrahim MN, Osman MK, Hussain Z. Comparison of MLP network training algorithms for fault classification in transmission lines. InControl System, Computing and Engineering (ICCSCE), 2014 IEEE International Conference on 2014 Nov 28 (pp. 208-213). IEEE.
26. Shahrukh Teli, Prashasti Kanikar. A Survey on Decision Tree Based Approaches in Data Mining.International Journal of Advanced Research in Computer Science and Software Engineering. 2015.
27. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 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. 1999 Jun 8;96(12):6745-50.
28. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine. 2001 Jun 1;7(6):673-9.
29. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science. 1999 Oct 15;286(5439):531-7.
30. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES. Gene expression correlates of clinical prostate cancer behavior. Cancer cell. 2002 Mar 31;1(2):203-9.
31. Srivastava B, Srivastava R, Jangid M. Filter vs. Wrapper approach for optimum gene selection of high dimensional gene expression dataset: An analysis with cancer datasets. InHigh Performance Computing and Applications (ICHPCA), 2014 International Conference on 2014 Dec 22 (pp. 1-6). IEEE.

Reviewers of this article

Prof.(Dr.) D.B. Wankhade

Professor & HOD, Department of ECE,
Rajiv Gandhi Institute of Technology,
Andheri (west), Mumbai -400 053, India

Prof. Y. Prapurna Chandra Rao

Assistant Proffessor, KLE University,
Belgaum, Karnataka

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.

We sincerely thank the above reviewers for peer reviewing the manuscript