DOIs:10.2015/IJIRMF/202302019 --:-- Research Paper / Article / Review

Two stage cascaded learner architecture for classification of multicategory microarray data

[1] **Dr. Sheela T,** [2] **Santhosh Kumar B N,** [3]**Prakasha Raje Urs M.**

[1]Associate Professor , Department of Computer Science, Maharani's Science College for Women, Mysore, INDIA.
[2]Assistant Professor , Department of Computer Science, Maharani's Science College for Women, Mysore, INDIA.
[3]Assistant Professor , Department of Computer Science, Maharani's Science College for Women, Mysore, INDIA.
Email - [1]sheela.mysore1.@gmail.com, [2]santhosh_bn@yahoo.com, [3]purs73@gmail.com

Abstract: *Cancer research is a major research area in medical field. The classification of different tumor types is very important and helpful in cancer diagnosis and drug discovery and it helps in providing better treatment. Microarray analysis is widely accepted for human cancer diagnosis and classification. In this paper, we have designed a two stage cascaded classification method, where two classifiers are in sequential order such that the first classifier is simple and the next classifier is more complex and it takes as input only the samples not classified correctly or a new sample not classified precisely by the previous classifier. Experiments conducted on multiclass microarray datasets show that the proposed cascading method is able to reach high prediction accuracies without increase in complexity and cost.*

Key Words: *cancer research, cascaded classification, cancer diagnosis, multiclass microarray data, ensemble classifier*.

1. INTRODUCTION:

Microarray is a useful technique for measuring expression data of thousands of genes simultaneously. The gene expression levels contain the fundamental information of the problems relating to the prevention and cure of diseases, drug discovery and biological evolution mechanism [1]. This technology is very useful in cancer classification. The purpose of classification on microarray data is to distinguish healthy tissue samples from cancerous tissue samples, as well as to further predict response to therapy. This kind of microarray data analysis is especially important in early tumour and cancer discovery because its result can effectively help cancer diagnosis and clinical treatment [2, 3]. Standard machine learning techniques cannot perform well for cancer classification, due to the characteristics of gene expression profiles such as high dimensionality and small sample size [4].

Efforts have been made to classify microarray data and to improve accuracy of classifiers [5, 6, 7]. Ensemble classification is a divide-and-conquer approach used to improve the performance. The idea of ensemble methodology is to build a predictive model by integrating multiple models to improve the prediction performance [8, 9, 10, 11, 12, 13].

In ensemble approach, a group of weak learners are combined to form a strong learner. Each learning algorithm finds a separate explanation for the data and converges to a different classifier and these classifiers make errors on different parts of the input space and they complement each other and an ensemble scheme can outperform the individual classifiers.

2. LITERATURE REVIEW:

The ensemble idea in supervised learning has been investigated since the late seventies. [14] suggests combining two linear regression models. Where, the first linear regression model is fitted to the original data and the second linear model to the residuals. There are two categories of ensemble techniques multi-expert systems and multistage systems. Multiexpert system is one in which different classifiers work in parallel and each one will give its own decision and final decision is made using a combiner. AdaBoost (Adaptive Boosting), a multi-expert system, was first introduced in [15], is a popular ensemble algorithm that improves the simple boosting algorithm via an iterative process. Some other examples for multiexpert systems are voting [16] mixture of experts [17] and stacked generalization [18]. Multistage system uses serial approach and is called cascading system. In this system the next classifier is consulted for sample classification only when the previous classifier is not confident on its decision and rejects the sample. By designing a multistage ensemble system with small number of classifiers, we can get good accuracy with less computation cost, memory usage and time. cascading uses the benefits of multistage properties and does not consult all classifiers on all instances and thus reduces classification time. Cascading, the multistage method of information fusion is discussed in

[19, 20]. Where, a sequence of classifiers are ordered under some conditions and the next classifier is only considered for patterns refused by the previous classifiers. The advantage of a cascading system is that, an earlier classifier handles major cases and a complex classifier in the next stage is only utilized with a small possibility hence not increasing the complexity greatly. A two stage classifier architecture for text document classification is used in [21] to automatically handle rejections, where, documents can be either classified or rejected at the first stage and the rejected documents are automatically classified at the second stage. A two-stage cascading scheme for iris recognition is presented by [22] and results prove that the cascading classification system outperforms single classifier. A registration method of cascading two fingerprint registration schemes is designed by [23] and a series of experiments validate the effectiveness of the multi-stage strategy. In [24], a decision boundary for the binary class datasets is obtained using the statistical method of confidence interval.

3. PROPOSED MODEL:

Based on the multi-stage architecture as explained in [19], we have selected a multistage system with only two learning algorithms and we call it a two stage cascading system. In the first stage we are using confidence interval method [24] of prediction of class labels and in the second stage we have used various classifiers like kNN (k-Nearest Neighbor Classifier), NBC (Nave Bayes Classifier) and DT (Decision Trees) and a comparison is made with individual classifier performance and also cascading with confidence interval prediction method.

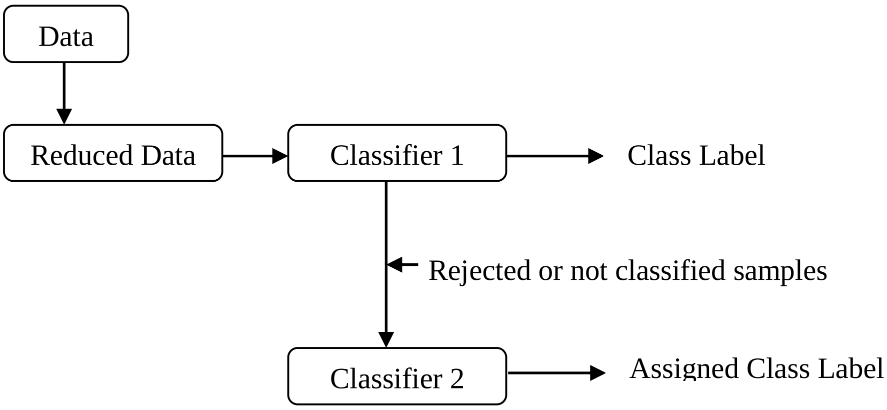

Figure 1: Proposed Cascaded Classifier Model

4. METHODOLOGY:

The first step is to normalize the microarray gene expression data and then reduce this pre-processed data into smaller subset. The performance and robustness of classification algorithms are improved when only a few features are involved in the classification. Thus, selecting relevant features for the construction of classifiers is important. In this paper Fast Correlation Based Filter (FCBF) [25] method is used to select significant genes. After gene selection, the proposed classification algorithm is applied to classify the reduced dataset.

FCBF gene selection method

Fast Correlation Based Feature selection is a filter model [25] which is designed for high-dimensional data and it effectively removes both irrelevant and redundant features and is less costly in computation than the currently available algorithms. FCBF works in an iterative way, where one good feature is selected at each step. A feature is said to be good if it is predominant in predicting the class concept and feature selection is a process that identifies all predominant features to the class concept and removes the rest. FCBF is used as a gene selection method with promising results in [26, 27].

Confidence Interval of genes

Our previously proposed method [24] works on the principle that statistically defined Confidence Interval can be used to obtain decision boundaries of the classes. For any parameter, confidence interval is an interval of numbers where, the true value of the population parameter is contained within this interval with certain specified accuracy. In this work, we have considered the level of confidence interval as 95%. It means that, when the experiment and the fitting procedure is repeated a number of times, 95% of the times the true parameter value will lie within this confidence interval. The algorithm for confidence interval class prediction is extended for multiclass microarray datasets as follows :

Let G=(g1, g2, g3, . . . gd) be the feature set of dataset D of size n and dimension d.

C={Cj} be the class set j=1,2,3... L

Confidence interval (CI_{ij}) of ith feature g_i for the jth class Cj is defined as

$$CI_{ij} = \left[g_{ij}^{-}, g_{ij}^{+}\right] = \left[\mu - \frac{t*\sigma}{\sqrt{n}}, \mu + \frac{t*\sigma}{\sqrt{n}}\right]$$

Where,

$g_{ij}^{-}, g_{ij}^{+}$are lower and upper bound of the confidence interval.

n=number of elements in the vector

μ_{ij} denotes mean of ith feature g_i for class Cj and

$\boldsymbol{\sigma}_{ij}$ denotes standard deviation of ith feature gi from its mean μ_{ij}

t is value obtained from t-distribution table for degree of freedom (n-1)

Algorithm

Step 1 : Data is normalized to have zero mean and unit standard deviation. This is done for each gene.

Step 2 : Find significant genes using FCBF selection method

Step 3 : For a test sample from class 1, apply confidence interval method of class prediction on the reduced dataset, as follows

a. $classct_j = 0$ (for j = 1,2,3 ... L)

b. Compute confidence interval $\left[g_{ij}^{-}, g_{ij}^{+}\right]$of all genes (for i = 1,2, . . . m) in class c_j (for j=2,3...L) for all samples.

c. Compute confidence interval $\left[g_{i1}^{-}, g_{i1}^{+}\right]$ of all genes (for i = 1,2, . . . m) in class c_1 excluding test sample.

d. Suppose that tg_i (for i=1,2 … m) are expression values of gene g_i of the test sample t ,

If $tg_i \in \left[g_{i1}^{-}, g_{i1}^{+}\right]$ then $classct_1 = classct_1 + 1$

If $tg_i \in \left[g_{ij}^{-}, g_{ij}^{+}\right]$then $classct_j = classct_j + 1$ (for j=2,3...L)

e. Assign class label k for the sample where $classct_k$ is max $\{classct_j\}$

Step 4 : Repeat step 3 for each of the samples in other classes c_j (for j=2,3...L).

Step 5 : Find the incorrectly classified samples in each class and samples classified in more than one class.

Step 6: Samples in step 5 are given as input to the next stage classifier.

The algorithm proposed can also be used to classify a sample whose label is not known. If this is uniquely classified in step 3 algorithm halts. Otherwise we use another classifier to assign a class label.

Classifiers used to cascade with the proposed confidence interval method.

k-Nearest Neighbor Classifier (kNN) : The k-NN method was first introduced by [28]. Here, in the first stage k nearest neighbours are determined and in the second stage class is determined using these neighbours. A test sample is given in vector form as input and the Euclidean distance between this and the vector representation of each training example is computed. The training sample closest to the test sample is considered as its Nearest Neighbor. And its class label is allocated to the test sample.

Nave Bayes Classifier (NBC): A Naïve Bayes Classifier (NBC) is a simple probabilistic classifier [29] based on Bayes theorem where every feature is assumed to be class-conditionally independent. In naïve Bayes learning, each instance is described by a set of features and takes a class value from a predefined set of values. When classifying an unknown sample, The NBC calculates the posterior probability of the sample belonging to each class, by comparing the distributions of the. samples features to those identified during training.

Decision Tree: Decision tree, also known as classification tree [30]. Leaves of the tree represents class labels and branches are combinations of features that lead towards a leaf. This classifier has internal nodes, based on which the tree splits into branches or edges. The branch that does not split is the leaf node. The leaf nodes are labelled with a single class label. The decision tree classification is a two-step process. In the first step, a decision tree is built from the training samples and the data is split using internal nodes into subsets with better class seperability. In the second step the tree is trimmed or pruned to introduce classification error on the test data.

5. DATASETS AND EXPERIMENTS:

Multiclass microarray gene expression datasets used for experimentation are given in Table 1. The number of samples and genes in the dataset and the number of classes present are also given. These datasets are available from http://www.gems-system.org

Table 1: Multiclass datasets

Dataset	Genes(m)	Samples(N)	Number of Classes
D1 Leukemia1 (ALL-B,T, AML)	5327	72	3
D2 Leukemia2(AML, ALL, MLL)	5327	72	3
D3 SRBCT (Small Round Blue Cell Tumor)	2308	83	4
D4 Brain_Tumor2	10367	50	4
D5 Brain_Tumor1	5920	90	5
D6 Lung cancer	12600	203	5

Experimental results on well-known gene expression datasets illustrate the effectiveness of the proposed approach. Standard classifiers such as k-NN, Naive Bayes (NB) and Decision Tree (DT) are used in addition to our proposed confidence interval class prediction method.

The following parameters are used for the classifiers in this work: k-Nearest neighbors (KNN) classification algorithm was applied with k = 1 and Euclidean distance metric is the distance measure. For Decision trees (DT) method, the minimum parent size (number of observations) is set as 10 and the minimum leaf size as 1. A kernel distribution is specified for predictors in the Naïve Bayes classification(NBC) algorithm. Since most microarray datasets only have relatively few samples, we chose the leave-one-out cross-validation method for evaluation. In this validation method, one of the samples is left as test sample and remaining samples are considered as training samples. The model is then used to predict the class label of the left-out sample. This is repeated for all samples.

Table 2. Classification accuracy with individual classifiers

Dataset	Performance of individual classifiers (%)			
	CI	kNN	NBC	DT
D1	31.94	91.67	93.05	87.5
D2	42.55	82.22	65.55	73.11
D3	61.11	97.59	93.46	83.13
D4	48.61	93.05	93.05	87.5
D5	35.11	81.11	90	91.11
D6	36.11	98.78	97.59	90.36

Table 3. Classification accuracy with cascaded classifier

Dataset	Performance of cascaded classifier		
	CI -> kNN	CI -> NBC	CI -> DT
D1	94.44	95.83	97.22

D2	88.89	71.11	84.44
D3	97.59	100	91.57
D4	94.44	95.83	97.22
D5	87.28	91.11	80
D6	100	97.59	92.77

Table 2. Shows the accuracy of the individual classifiers for multiclass datasets. Table 3. Shows the accuracy with our proposed cascaded method. It is evident from Table 3, that for many datasets close to 40% of the instances are processed in the first stage of the cascade, reducing the complexity by not using the second stage classifier.

6. FINDINGS:

For dataset D1, the classifiers kNN, NBC and DT cascaded with the proposed confidence interval method of class prediction shows an improvement of 3%, 2% and 10% in accuracy compared to individual classifiers. For dataset D2, there is an improvement of 6%, 6% and 11%. Similarly for all the other datasets, cascaded method is giving an increase of more than 2% in accuracy compared to individual classifiers. For dataset D3 and D6 cascaded method with NBC and kNN gives 100% accuracy respectively.

7. CONCLUSION:

In this work we have presented a cascade classification method, in which a less complex confidence interval class prediction method is used in the first stage and more complex classifier in the next stage. The idea is to cascade a small number of classifiers sequenced in terms of complexity of the classification algorithm and thus later classifier is used unless actually necessary. The experimental results show that our proposed method is effective and efficient in predicting normal and tumour samples. Confidence Interval method of class prediction can be easily incorporated into the popular classifiers without having to change the underlying algorithms.

REFERENCES:

1. Basford2013] K.E.Basford *et al*.,"On the classification of microarray gene-expression data". Briefings in Bioinformatics, 14(4), 402–410, 2013.
2. Dupuy2007 A.Dupuy and R.Simon," Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting", J Natl Cancer Inst ;9:147–57, 2007.
3. Boulesteix 2008 A.L.Boulesteix *et al*., "Evaluating microarray-based classifiers: an overview", Cancer Inform ; 6:77–97, 2008.
4. T.R.Golub *et al*.," Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", SCIENCE(1999), Vol 286,531–537, 1999.
5. Ahmed, O., and Brifcani, A. (2019, April). *Gene Expression Classification Based on Deep Learning*. 4th Scientific International Conference Najaf (SICN) pp. 145-149, 2019.
6. Cahyaningrum, K., and Astuti, W. *Microarray Gene Expression Classification for Cancer Detection using Artificial Neural Networks and Genetic Algorithm Hybrid Intelligence*. International Conference on Data Science and Its Applications (ICoDSA) (pp. 1-7). IEEE, 2020.
7. Hatim Z Almarzouki. *Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile*. Journal of Healthcare Engineering, Article ID 4715998, 13 pages, https://doi.org/10.1155/2022/4715998, 2022.
8. Dietterich TG2000 Dietterich TG. Ensemble methods in machine learning. In: Proceedings of Multiple Classifier System.vol. 1857.Springer; 2000. pp. 1–15.
9. Saeys Y, Thomas Abeel, Yves Van de Peer. :Robust feature selection using ensemble feature selection techniques. In Proceedings of the 25th European Conference on Machine Learning and Knowledge Discovery in Databases, Part II, Springer-Verlag, Berlin, Heidelberg, pp. 313–325 (2008).
10. Lai C. M., and Huang H. P. *A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique*. Applied Soft Computing, 106994, 2020.
11. Maniruzzaman M, et al. *Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms*. Comput Methods Prog Biomed;176:173–93, 2019.
12. Othman M.S., Kumaran S. R., and Yusuf L.M. *Gene Selection Using Hybrid Multi-Objective Cuckoo Search Algorithm with Evolutionary Operators for Cancer Microarray Data*. IEEE Access, 8, 186348-186361, 2020.
13. Zhang X., He T., Ouyang L., Xu X., and Chen S. *A Survey of Gene Selection and Classification Techniques Based on Cancer Microarray Data Analysis*. IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 1809-1813) IEEE, 2018.

14. Tukey JW (1977); Exploratory data analysis. Addison-wesley series in behavioral science, First Edition.
15. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the 13th international conference, pp325-332
16. Kittler 1998 Kittler, J., Hatef, M. Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 226-239.
17. Jocobs 1991 Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991) Adaptive mixtures of local experts. *Neural Computation*, *3*, 79-87.
18. Wolpert, D. H. (1992) Stacked generalization. *Neural Networks*, *5*, 241-259.
19. Pudil P 1992 P. Pudil, J. Novovicova, S.Blaha and J. Kittler. Multistage Pattern Recognition with Rejection Option. Proceedings of the 11th International Conference on Pattern Recognition, Vol.B, pp. 92 - 95, 1992.
20. Kaynak 2000 C. Kaynak and E. Alpaydin. MultiStage Cascading of Multiple Classifiers: One Man's Noise is Another Man's Data. Proc. 17th International Conf. on Machine Learning, 2000.
21. Fumera, G., Pillai, I., & Roli, F. (2004). A Two-Stage Classifier with Reject Option for Text Categorisation. In *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 771–779). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-27868-9_84. Z
22. henan Sun 2004] Zhenan Sun, Yunhong Wang, Tieniu Tan and Jiali Cui. Cascading Statistical And Structural Classifiers For Iris Recognition. Proceedings of IEEE International Conference on Image Processing, 2004, pp.1261 - 1264.
23. Jin Qi, Zhongchao shi, Xuying Zhao and Yangsheng Wang. Cascading a Couple of Registration Methods for a High Accurate Fingerprint Verification System. Proceedings of Sinobiometrics'04, LNCS 3338, Beijing, China, Dec. 2004.
24. Sheela T, Lalitha Rangarajan, "Statistical Class Prediction Method for Efficient Microarray Gene Expression Data Sample Classification" , ICACCI2017,IEEE Explore, DOI: 10.1109/ICACCI.2017.8125819
25. Yu L., Liu H.. Feature selection for high dimensional data: a fast correlation-based filter solution, Proceedingd of teh Twentieth International Conference on Machine Learning(ICML-2003), 2003(pg.856-863)
26. Alireza Osareh and Bita Shadgar, "An Efficient Ensemble learning Method for Gene Microarray Classification", BioMed Research International, Volume 2013, Article ID 478410.
27. Djellali, H., Guessoum, S., Ghoualmi-Zine, N., & Layachi, S. (2017). Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection. In *2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B)* (pp. 1–6). https://doi.org/10.1109/ICEE-B.2017.8192090.
28. E.Fix and J.L.Hodges," Discriminatory Analysis-Nonparametric Discrimination :Consistency Properties. Techical Report, 21-49-004. Report no.4 US Air Force School of Aviation Medicine, Randolph Field, 261-279, 1951.
29. Friedman JH, Bentley JL, Finkel RA. An Algorithm for Finding Best Matches in Logarithmic Expected Time. ACM Trans.Math. Softw. 1977 sep; 3(3):209–226.
30. Quinlan JR. Simplifying decision trees. International Journal of Human-Computer Studies 1999; 51(2):497.