ISSN 2047-3338

A Comparative Study on Gene Selection and Classification Methods for the Cancer Subtypes Prediction

Pheba Thomas
Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, India
phebs2007@gmail.com

***Abstract*— Microarray gene expression data gained great importance in recent years due to its role in disease diagnoses and prognoses which help to choose the appropriate treatment plan for patients. Interpreting gene expression data remains a difficult problem and an active research area due to their native nature of high dimensional low sample size. These issues poses great challenges to existing classification methods. Thus effective feature selection techniques are often needed in this case to aid to correctly classify different tumor types and consequently lead to improve treatment strategies. Small sample size remains a bottleneck to design suitable classifiers. Traditional supervised classifiers can only work with labeled data. On the other hand, a large number of microarray data that do not have adequate follow-up information are disregarded. Particular, the study report focus on the most used data mining techniques for gene selection and semi supervised cancer classification. In addition, it provides a general idea for future improvement in this field.**

***Index Terms*– KFRS, Microarray Data, TSVM, Unlabeled Samples and Gene selection**

I. INTRODUCTION

DEVELOPING simple data mining tests that allow early cancer detection is one of the top priorities in cancer research field. Cancer classification of different tumor types is of great importance in cancer diagnosis and drug discovery. Such tests will impact patient care and outcome through disease screening and early detection. Large number of gene expression/miRNA data and their diverse expression patterns indicate that they are likely to be involved in a broad spectrum of human diseases [1]. The advent of microarray technology has made it possible to study the expression profiles of a large number of genes across different experimental conditions. Microarray- based gene expression profiling has shown great potential in the prediction of different cancer subtypes [4], [5], [6], [7], [8]–[12].

Major research on extending support vector machines (SVMs) to handle semi labeled data is based on the following idea: solve the standard inductive SVM (ISVM) while treating the unknown labels as additional optimization variables. By maximizing the margin in the presence of unlabeled samples, one can learn the decision boundary that traverses through low-density regions while respecting labels in the input space. In other words, this approach implements the cluster assumption for semisupervised learning, that samples in a data cluster have identical labels. The idea was first introduced under the name of transductive SVM, but since it learns an inductive rule defined over the entire input space, the approach is referred to as semisupervised SVM (S3VM). Each cluster of samples is assumed to belong to one data class. Thus, a decision boundary is defined between clusters. A variety of semisupervised techniques have been proposed and many successful algorithms directly or indirectly assume high density within class and low density between classes, and can fail when the classes are strongly overlapping. This can be illustrated by comparing the well-known SVMs to their semisupervised extension, transductive SVM, progressive TSVM algorithm (PTSVM), transductive SVMs (TSVMs) and semisupervised SVMs (S3VMs). TSVMs and S3VMs are iterative algorithms that use SVMs to gradually search a reliable hyper plane exploiting both labeled and unlabeled samples in the training phase [13].

By using supervised classification method, only labeled data can be used but it is very expensive and difficult to obtain. This paper mainly concentrated on semi supervised classification method which uses both labeled and unlabeled data [8]. Several semi supervised classification methods discussed are transductive support vector machine, recursively partition model, cut edge and nearest neighbour rule, low density separation approach.

One major problem faced in the classification is due to large number of features of dataset. If there are thousands of features then it will affect the performance of classifier. This is one of the challenges in machine learning technique and is called feature selection [1], [7]. Selection of informative genes is an important part for the analysis of microarray data. Successful feature selection has several advantages in such situations where thousands of features are involved. First, dimension reduction is employed to reduce the computational cost. Second, reduction of noises is performed to improve classification accuracy. Finally, extraction of more interpretable features or characteristics that can be helpful to identify and monitor the target diseases.

In this work, we have investigated several gene selection methods namely kernelized fuzzy rough set (KFRS), Fuzzy Rough Set Attribute Reduction on Information Gain

Ratio(FRS_GR),signal-to-noise ratio (SNR), positive approximation, Fuzzy entropy measure feature selection with similarity classifier, positive approximation based on rough set theory and consistency based feature selection (CBFS) .

II. SEMISUPERVISED CLASSIFICATION TECHNIQUES

Semi-supervised learning is a learning paradigm concerned with the study of how computers and natural systems such as humans learn in the presence of both labeled and unlabeled data.

A) Inductive SVM

[24] Inductive SVM (ISVM) is a general class of learning architecture originated in modern statistical learning theory.Given a training dataset, the SVM training algorithm obtains the optimal separating hyperplane in terms of generalization error. In a binary classification problem, let S = [(xi; yi)], i=1,2,.....,l be the set of training examples, where yi $\in \{\pm 1\}$ is the label associated with input pattern xi. In a learning problem, the task is to estimate a function f from a given class of functions that correctly classifies unseen examples (x,y) by computing the sign(f (x)). In the case of pattern recognition, this means that given some new patterns x$\in \chi$, the classier predicts the corresponding y$\in \{\pm 1\}$.

B) Graph Based Semi Supervised Learning

[26] Another method of semi-supervised learning based on a Gaussian random field model. Labeled and unlabeled data are represented as vertices in a weighted graph, with edge weights encoding the similarity between instances. The model will provide effective structure of unlabeled data so it helps to improve classification accuracy. Gaussian random field model has concentrated on the use of only the mean of the field, which is characterized in terms of harmonic functions and spectral graph theory.

C) Progressive Transductive SVM

[15] By taking a transductive approach instead of an inductive one in support vector classifiers, the working set can be used as an additional source of information about margins. Compared with traditional inductive support vector machines, transductive support vector machine is often more powerful and can give better performance. In transduction, one estimates the classification function at points within the working set using information from both the training and the working set data. This will help to improve the generalization performance of SVMs, especially when training data is inadequate. Intuitively, we would expect transductive learning to yield improvements when the training sets are small or when there is a significant deviation between the training and working set subsamples of the total population. A progressive transductive support vector machine is addressed to extend Joachims transductive SVM to handle different class distributions. PTSVM can automatically adapt to different data distributions and realize a transductive learning of support vectors in a more general sense. It solves the problem of having to estimate the ratio of positive/negative examples from the working set. This algorithm is very promising.

D) Low Density Separation Approach

[16] The cluster assumption is key to successful semi-supervised learning. Based on this, we propose three semi-supervised algorithms: Deriving graph-based distances that emphasize low density regions between clusters, followed by training a standard SVM, optimizing the transductive SVM objective function, which places the decision boundary in low density regions, by gradient descent, combining the first two to make maximum use of the cluster assumption.

E) Large Margin Classification

[25] A large margin semisupervised learning method, which aims to extract the information from unlabeled data for estimating the Bayes decision boundary. This is achieved by constructing an efficient loss for unlabeled data with regard to reconstruction of the Bayes decision boundary. This method is using both the grouping (clustering) structure of unlabeled data, is designed to recover the classification performance based on complete data without missing labels. One of the applications is to predict the gene function. This method integrates labeled and unlabeled data through using the clustering structure of unlabeled data. One critical issue is how to use unlabeled data to enhance the accuracy of estimating the generalization error so that adaptive tuning is possible.

F) Nearest Neighbour Rule and Cut Edges

[17] First step of this approach, a relative neighborhood graph based on all training samples is constructed for each unlabeled sample, and the unlabeled samples whose edges are all connected to training samples from the same class are labeled. These newly labeled samples are then added into the training samples. In the second step, standard self-training algorithm using nearest neighbor rule is applied for classification until a predetermined stopping criterion is met. In the third step, a statistical test is applied for label modification, and in the last step, the remaining unlabeled samples are classified using standard nearest neighbor rule.

G) Semi-Supervised SVMs

[27] The main goal of semi-supervised learning is to employ the large collection of unlabeled data jointly with a few labeled examples for improving generalization performance. Semi-Supervised Support Vector Machines are based on applying the margin maximization principle to both labeled and unlabeled data. Unlike SVMs, their formulation leads to a non-convex optimization problem. A suite of algorithms have recently been proposed for solving S3VMs.

H) Transductive SVM

[13] To overcome the limitation of small sample size TSVMs are used which are basically iterative algorithms that gradually search the optimal separating hyperplane in the feature space with a transductive process that incorporates unlabeled samples in the training phase. Unlike the selection procedure of transductive samples in traditional TSVMs, the selection of transductive samples is done through a process of filtering the unlabeled samples. Unlabeled samples that fall into the margin will have richer information to find a better separating hyper plane these samples are selected and working set is updated and retrained the TSVM for each iteration. This procedure improves the generalization capability of the classifier. Gradually, the separating hyperplane will move to a finer position in subsequent iterations.

III. GENE SELECTION METHODS

Feature selection is an effective technique in dealing with dimensionality reduction. For classification, it is used to find an "optimal" subset of relevant features such that the overall accuracy of classification is increased while the data size is reduced and the comprehensibility is improved. Feature selection methods contain two important aspects: evaluation of a candidate feature subset and search through the feature space.

A) Consistency-Based Feature Selection

[21] Inconsistency measure according to which a feature subset is inconsistent if there exist at least two instances with same feature values but with different class labels. We compare inconsistency measure with other measures and study different search strategies such as exhaustive, complete, heuristic and random search that can be applied to this measure. Consistency measure with other measures shows that it is monotonic, fast, multivariate, capable of handling some noise and can be used to remove redundant and/or irrelevant features.

B) A Signal-to-Noise Ratio Approach

[2] Signal to noise ratio ranking is another model for feature selection it proposes two approaches. In first approach, the genes of microarray data is clustered by k-means clustering and then SNR ranking is implemented to get top ranked features from each cluster and given to two classifiers for validation such as SVM and k-NN. In the second approach the features of microarray data set is ranked by implementing only SNR ranking and top scored feature are given to the classifier and validated. First approach for feature selection is better than to second approach as due to clustering technique similar features will be grouped in to the same clusters. After applying SNR ranking and selecting top scored features from each cluster may give a true pattern which helps to enhance the classification accuracy. But in case of second approach after applying SNR ranking we can randomly choose the top scored features where we can get redundant feature or noisy features with similar SNR score and does not provide any relevant information about the data.

C) Rough and Fuzzy-Rough-Based Approaches

[22] Semantics-preserving dimensionality reduction refers to the problem of selecting those input features that are most predictive of a given outcome; a problem encountered in many areas such as machine learning, pattern recognition, and signal processing. This has found successful application in tasks that involve data sets containing huge numbers of features (in the order of tens of thousands), which would be impossible to process further. There are two approaches for semantics-preserving dimensionality reduction rough and fuzzy-rough-based approaches. Rough seection include rough set attribute reduction, reduction with variable precision rough sets dynamic reducts. Fuzzy rough attribute reduction includes fuzzy equivalence classes fuzzy lower and upper approximations fuzzy-rough reduction process, reduct computation, rough Set-Based Feature Grouping. Conventional rough set methods are unable to deal with real-valued attributes effectively. This prompted research into the use of fuzzy-rough sets for feature selection.

D) Positive Approximation

[18] Existing heuristic attribute reduction has several limitations. Yuhua Qian proposed a method called positive approximation based on rough set theory [6]. The main objective of this method is to select some property of original data without any redundancy. There will be more than one reduct. But only one reduced attribute is needed so a heuristic algorithm is proposed based on significance measure of the attribute. This method is somewhat similar to greedy search algorithm. But some modification is proposed here significance measure is calculated. Until the reduct set is obtained the attribute with maximum significance value is added at each stage. The result of positive approximation of attribute reduction shows that it is an effective accelerator. There are three speedup factors in positive approximation based feature selection:

- One attribute can select more than one in each loop. So this will helps to provide a restriction in the result of the reduction algorithm.
- Reduced computational time due to attribute significance measure.
- Another important factor in this algorithm is size of the data is reduced and time taken for the computation of stopping criteria is also reduced to minimum.

E) Fuzzy Entropy Measures with Similarity Classifier

[14] Using fuzzy entropy-based feature selection combined with similarity classifier, we managed to reduce the computational time and simplify the data set by using only subset of features instead of the whole data set to do the classification. Feature selection method using fuzzy entropy

measures together with similarity classifier is giving good result.

F) Fuzzy Preference Rough Set

[14] Pawlak's rough set model is constructed based on equivalence relations. The relations have one of the main limitations when applying this model to complex decision tasks. On the other hand, fuzzy preference relations can reflect the degree of preference quantitatively making it more powerful in extracting information from fuzzy data than equivalence or dominance relations. This motivates to use FPRS technique for gene selection.

G) Kernelized Fuzzy Rough Set for Feature Selection

[24] High level of similarity between kernel methods and rough sets can be obtained using kernel matrix as a relation [19].Kernel matrices could serve as fuzzy relation matrices in fuzzy rough sets. Taking this into account, a bridge between rough sets and kernel methods with the relational matrices was formed [19]. Kernel functions are used to derive fuzzy relations for rough sets based data analysis. In this study, Gaussian kernel approximation has been used to construct a fuzzy rough set model, where sample spaces are granulated into fuzzy information granules in terms of fuzzy T - equivalence relations computed with Gaussian kernel.

Table 1: Comparison of different Gene selection techniques and Classification methods

Method	Merits	Demerits
Large margin semisupervised learning method	• Unlabeled and labeled datasets can be used. • Maximum margin classification gives better accuracy for prediction	Feature selection is not performed
Semi-Supervised Support vector Machines	• Less error • use large collection of labeled data • Margin maximization will give better accuracy	Performance is less because unlabeled data is used in very small percent.
Nearest neighbor rule and cut edges	better classification performance.	computational complexity is high
Low Density Separation approach	Better accuracy	large sets of unlabeled data cannot be used
Progressive transductive support vector machine	Very efficient in low dimensional dataset	• Feature selection is not performed • Not good for gene expression dataset
Graph based semi supervised learning	• High classification Accuracy when the labeled data are few. • Outperforms when the training data are sufficient • Good for supervised case	• computational complexity is high • Not suitable for unlabeled dataset
TSVM	• Accuracy of classification is increased by using both labeled and unlabeled data • Good for gene expression data	have to estimate the no: of positive/negative examples
CBFS	• Accuracy of classification is increased • data size is reduced	computational time is more
Rough set theory	• Performance is better with large dataset • Less time consuming	Cannot use unlabelled data
Rough and fuzzy-rough-based approaches	Less costly	Time consuming
Fuzzy-Rough Techniques	Hybrid data reduction	Costly

IV. CONCLUSION AND FUTURE WORK

We presented a comparative study of state-of-the-art gene selection methods and classification methods based on gene expression data. The efficiency of eight semisupervised classification methods and seven different gene selection methods was compared. The merits and demerits of these methods are tabulated.

In the future, we plan to study the effect of doing feature selection and classification altogether. One possible way is to use wrapper methods. Another possible way is to combine the feature selection in the classification learning process. For instance, we can add a feature selection term in the objective function of svms, so that the optimization problem can do feature selection and classification simultaneously. Another direction of future research is to combine the information from different data sets together. It is a commonly seen scenario that there are a number of biological data sets that share the same features but are collected by different groups under different experimental conditions. As a scope of further development, several issues remain open to be addressed: 1) integration of other sources of information could be important to enhance clinical/translational research. 2) different combination of feature selection methods needs to be investigated to obtain more biologically relevant genetic signatures and 3) the concept of fuzzy set theory could be introduced in semisupervised learning to improve model development.

REFERENCES

[1] E. Berezikov, E. Cuppen, and R.H.A. Plasterk, "Approaches to microRNA discovery," Nature Genet., Vol. 38, pp. S2S7, May 2006.

[2] Barnali Sahu, Debahuti Mishra (2011), Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach, International Journal of Scientific & Engineering Research, Vol. 2, Issue 4, ISSN 2229-5518.

[3] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, Sep 1994.

[4] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," Bioinformatics, Vol. 23, No. 21, pp. 2859–2865, 2007.

[5] S. Bandyopadhyay, R. Mitra, and U. Maulik, "Development of the human cancer microRNA network," BMC Silence, Vol. 1, No. 6, 2010.

[6] A.J. Gentles, S.K. Plevritis, R. Majeti, and A. A. Alizadeh, "Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia," JAMA—J. Amer. Med. Assoc., Vol. 304, No. 24, pp. 2706–2715, 2010.

[7] H. K. Kim, I. J. Choi, C. G. Kim, A. Oshima, and J. E. Green, "Gene expression signatures to predict the response of gastric cancer to cisplatin and fluorouracil," J. Clin. Oncol., Vol. 27, No. 15s, 2009.

[8] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes," BMC Bioinformat., Vol. 10, No. 27, 2009.

[9] U.Maulik, "Analysis of gene microarray data in soft computing framework," Appl. Soft Comput., vol. 11, no. 6, pp. 4152–4160, 2011.

[10] A. Mukhopadhyay, S. Bandyopadhyay, and U. Maulik, "Multi-class clustering of cancer subtypes through SVM based ensemble of pareto-optimal solutions for gene marker identification," PLoS ONE, Vol. 5, No. 11, pp. 1–14, 2010.

[11] U.Maulik,S.Bandyopadhyay,andA.Mukhopadhyay,Multiobject ive Genetic Algorithms for Clustering: Applications Data Mining and Bioinformatics. New York: Springer-Verlag, 2011.

[12] U. Maulik and A. Mukhopadhyay, "Simulated annealing based automatic fuzzy clustering combined with ANN classification for analyzingmicroarray data," Comput. Oper. Res., vol. 37, no. 8, pp. 1369–1380, 2010.

[13] U. Maulik, A. Mukhopadhyay, and D. Chakraborty, "Gene-expression based cancer subtypes prediction through feature selection and transductive SVM," IEEE Trans. Biomed. Eng., vol. 60, no. 4, pp.1111–1117, 2013.

[14] U. Maulik and D. Chakraborty, "Fuzzy Preference Based Feature Selection and Semisupervised SVM for Cancer Classification," IEEE Transactions on nanobioscience, vol. 13, no. 2, June 2014.

[15] Chen.Y, Dong.S and Wang.G (2003), "Learning with progressive transductive support vector machine," Pattern Recognit. Lett., Vol. 34, No. 12, pp. 1845–1855.

[16] Chapelle.O and Zien.A (2005), "Semi-supervised classification by low-density separation," in Proc. 10th Int. Works. Artif. Intell. Stat., pp. 57–64.

[17] Yu Wang , Xiaoyan Xu, Haifeng Zhao , Zhongsheng Hua "semi-supervised learning based on nearest neighbor rule and cut edges" Knowledge-Based Systems 23 (2010) 547– 554

[18] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation:An accelerator for attribute reduction in rough set theory," Artif. Intell., Vol. 174, pp. 597–618, 2010.

[19] Q. Hu, D. Yu, W. Pedrycz, and D. Chen, ``Kernelized fuzzy rough sets and their applications," IEEE Trans. Knowl. Data Eng., vol. 23, no. 11,pp. 16491667, Nov. 2011.

[20] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu. Gaussian Kernel Based Fuzzy Rough Sets: Model, UncertaintyMeasuresandApplications.[Online].Available:http://www4.comp.polyu.edu.hk/

[21] M. Dash and H. Liu, "Consistency-based search in feature selection," Artif. Intell., Vol. 151, No. 12, pp. 155176, Dec. 2003.

[22] Jensen. R and Shen. Q (2004), "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches," IEEE Trans. Knowl. DataEng., Vol. 16, No. 12, pp. 1457–1471.

[23] V. N. Vapnik, Statistical Learning Theory. New York, NY, USA: Wiley, 1998.

[24] U. Maulik and D. Chakraborty, "Identifying Cancer Biomarkers From Microarray Data Using Feature selection and Semisupervised learning," IEEE Journal of Transactions in Health and Medicine, Vol. 13, No. 2, Dec 2014.

[25] Shen. X.T, Wang. J.H (2007), "Large margin semi-supervised learning," J. Mach. Learning Res., vol. 8, pp. 1867–1891.

[26] Ghahramani. Z, Lafferty. J and Zhu. X (2003), "Semi supervised learning uses Gaussian fields and harmonic functions," in Proc. 20th Int. Conf. Mach. Learning.

[27] Chapelle. O, Keerthi. S.S and Sindhwani. V (2008), "Optimization techniques for semi-supervised support vectors," J. Mach. Learn. Res., Vol. 9, pp. 203–233.