A Survey on Different Feature Selection Method for Cancer Biomarker Discovery.

G. Keerthana
M.E-CSE(1st Year)
K.S.Rangasamy College of Technology
Tiruchengode, India

Dr. K. Sakthivel
Professor
K.S.Rangasamy College of Technology
Tiruchengode,India

***Abstract*—Image processing in medical diagnosis involve stages such as image capture, image enhancement, image segmentation and feature extraction Identifying key biomarkers for different cancer types can improve diagnosis accuracy and treatment. The identification of biomarkers for early detection could be a promising strategy to decrease mortality. Gene expression data can help differentiate between cancer subtypes. However the limitation of having a small number of samples versus a larger number of genes represented in a dataset leads to the overfitting of classification models. Feature selection is a dimensionality reduction technique widely used and is one of the key topics in machine learning and other related fields it can remove the irrelevant even noisy features and hence improve the quality of the data set and the performance of learning systems. Feature selection methods can help select the most distinguishing feature sets for classifying different cancers. Many feature selection approaches like F-statistic, Maximum Relevance Binary Particle Swarm Optimization (MRBPSO) and Class Dependent Multi- category Classification (CDMC), Biomarker Identifier system are available. This feature selection method combines filter and wrapper based methods. This method improve the computation efficiency and are robust against overfitting**

***Keywords*— Features, Biomarkers, Classification, Filter, Wrapper**

I. INTRODUCTION

Classification of data has been successfully applied to a wide range of application areas, such as scientific experiments, medical diagnosis, credit approval, weather prediction, customer segmentation, target marketing and fraud detection. The development of gene expression technologies such as microarray and RNAseq has made it easier to monitor the expression pattern of thousands of genes simultaneously and a huge amount of gene expression data has been produced during these experiments. Cancer classification has been investigated for the identification of tumour biomarkers computationally.

Expression datasets normally consist of a large number of genes compared with a limited number of samples. Due to the high dimensionality of gene expression data, feature selection techniques are used to select a small subset of key genes that change under different cancer conditions. This potentially decreases clinical cost by testing on fewer biomarker genes and improves the accuracy of disease diagnosis by reducing data dimensionality and removing noisy features.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.

A feature selection technique is to be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples.

Feature selection methods are broadly divided into filter, wrapper and embedded methods. Filter type methods select variables regardless of the model. They are based only on general features like the correlation with the variable to predict. Filter methods suppress the least interesting variables. The others variables will be part of the model classification, a regression used to classify or a data prediction. These methods are particularly effective in computation time and robust to overfitting. Wrapper Method evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. The two main disadvantages of these methods are:

- The increasing overfitting risk when the number of observations is insufficient.
- The significant computation time when the number of variables is large.

Embedded methods have been proposed to reduce the classification of learning. They try to combine the advantages of both previous methods. The learning algorithm takes advantage of its own variable selection algorithm. So, it needs to know preliminary what a good selection is, which limits their exploitation. Many filter and wrapper based methods have been applied for feature selection such as tabu search, random forest , mutual information, entropy based method, regularized least square, and support vector machine

II. FEATURE EXTRACTION

Feature extraction starts from an initial set of measured data and builds derived values intended to be informative, non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features. This process is called as Feature extraction. The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

III. CONVENTIONAL METHODS

Some of the conventional methods used are described below.

3.1 Support Vector Machine

The Support Vector Machine (SVM) is a statistical learning method first proposed by N.Vapnik in 1963. It is based on the theories of VC dimension and structure risk minimization. For two-class classification problems, SVM uses a nonlinear mapping known as a kernel function to map the training data into a higher dimensional feature space, and construct an optimal separating hyperplane in the higher dimensional space corresponding to a nonlinear classifier in the input space. With the kernel functions and the high dimensional space, the hyperplane computation requires solving a quadratic programming problem using Lagrange multipliers:

$$\underset{\mathbf{w},\xi,b}{\arg\min}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right\}$$

Subject to (for any i =1,…,n)

$$y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

C is a tuning parameter that allows the user to control the tradeoff between classifying the training samples without error and maximizing the margin. Instead of solving this primal problem, it is always a practice to solve its dual problem.

$$\tilde{L}(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$0 \leq \alpha_i \leq C, \text{ and}$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0.$$

αi denotes the Lagrange variable for the ith constraint. K (xi,xj) is the kernel function. This work uses the Radial Basis Function (RBF) kernel. The RBF kernel with a parameter γ (gamma):

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2)$$

3.2 Entropy-Based

By this method, the feature those have relatively random expression distribution can be filter out. The remaining features is found by finding some cut points in these features automatically. The value of the features ranges such that the resulting expression intervals of every feature can be distinguished maximally .If a feature containing the same class of sample induced by the cut point to every expression interval, then the cut point of this feature have some partitioning that have an entropy value of zero in an ideal case. Features have smaller entropy then it is more discriminatory. For considering those features with lowest entropy values sort the values of the entropy in ascending order.

3.3 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 same distribution for all trees in the forest. The generalization error for forests converges as to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. 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. A random forest is a classifier consisting of a collection of tree structured classifiers {h(**x**, Θk), k=1,...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x** .

Given an ensemble of classifiers $h_1(\mathbf{x})$, $h_2(\mathbf{x})$, ... , $h_k(\mathbf{x})$, and with the training set drawn at random from the distribution of the random vector Y,**X**, define the margin function as

$$mg(\mathbf{X},Y) = av_k\, I(h_k(\mathbf{X})=Y) - \max_{j\neq Y} av_k\, I(h_k(\mathbf{X})= j)$$

where *I*(**.**) is the indicator function. The margin measures the extent to which the average number of votes at **X**,*Y* for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. The generalization error is given by

$$PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X},Y) < 0)$$

where the subscripts **X**,*Y* indicate that the probability is over the **X**,*Y* space. In random forests,

$$h_k(\mathbf{X}) = h(\mathbf{X}, \ _k) .$$

3.4 Filter Based Methods

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Common measures include the mutual information, the pointwise mutual information, Pearson product-moment correlation coefficient, inter/intra class distance or the scores of significance tests for each class/feature combinations. Filters are usually less computationally intensive than wrappers, but they produce a feature set which is not tuned to a specific type of predictive model. This lack of tuning means a feature set from a filter is more general than the set from a wrapper, usually giving lower prediction performance than a wrapper. However the feature set doesn't contain the assumptions of a prediction model, and so is more useful for exposing the relationships between the features. Many filters provide a feature ranking rather than an explicit best feature subset, and the cut-off point in the ranking is chosen via cross-validation. Filter methods have also been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems.

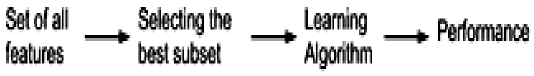

Fig 1: Filter method for feature selection.

3.4.1 Feature pre-selection with F-statistic

Thousands of genes are measured in each dataset compared to only tens of available samples. Generally, only a small number of changes in gene expression are related to each cancer type. Therefore, feature selection is an essential step to select a subset of significantly changed genes for further classification of cancer samples. In this pre-selection stage, we use a filter method based on the F-statistic to select the top ranking genes for each dataset. These genes will be used for selecting feature subsets for each class in the next section. The F-test value for gene g in k classes is calculated in the following formula:

$$F(g_i) = \frac{MS_{among}}{MS_{within}} = \frac{SS_{among}}{DF_{among}} \Big/ \frac{SS_{within}}{DF_{within}}$$

where MS_{among} and MS_{within} represent the mean squares among and within groups(classes). SS_{among} and SS_{within} are sum of squares among and within groups of samples. DF stands for degree of freedom. The sum of squares can be computed in the following formula:

$$SS_{among} = \sum_{h=1}^{k} \sum_{j=1}^{N_h} (\bar{g}_h - \bar{g})^2$$

$$SS_{within} = \sum_{h=1}^{k} \sum_{j=1}^{N_h} (g_{hj} - \bar{g}_h)^2$$

where $\overline{gh}$ and $\overline{g}$ are the mean expression values for gene g in class h and overall k classes respectively. N_h is the number of samples for gene g within class h.

3.4.2 Biomarker Identifier

It is a filter-based feature selection method adopted to analyse gene expression data that might be used to discriminate between samples with and without cancer. BMI method combines various statistical measures to discern the ability of features to distinguish between two data groups of interest. It considers three measures for evaluating features. First it checks whether distribution of a feature is significantly different between data groups. If the distribution of a feature changes substantially, the feature might be relevant to the underlying difference between data groups. Second, the ratio of overall variance relative to variance in control group is used to measure the reliability of a feature. BMI penalizes or credits a score of a feature by the ratio of overall variance relative to variance in control group. Lastly BMI considers the discriminative power of each individual feature by incorporating the true positive rate from logistic regression using the feature.

3.5 Wrapper Based Methods

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

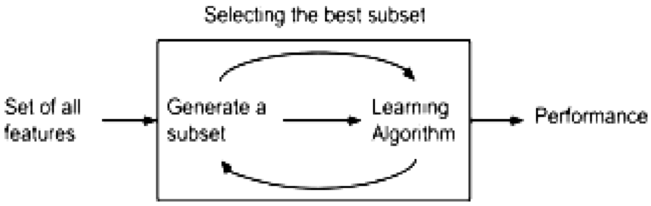

Fig 2: Wrapper Method for feature selection.

3.5.1 Binary Particle Swarm Optimization (BPSO)

The particle swarm optimization is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to given measure of quality. A binary version of particle swarm optimization (BPSO) was proposed for dealing with optimization problems with discrete and binary variables which cannot be handled well by PSO. In BPSO, an initial population of particles is generated with random positions and velocities. The position of particle i, Pi = (P_{i1}, P_{i2}, …, P_{in}), represents a potential solution of the optimization problem with n dimension. P_{ij} is a binary value of either 1 or 0 which represents if the corresponding feature is selected or not. The velocity is represented by V_i = (V_{i1}, V_{i2},…, V_{in}). V_{ij} represents the probability of bit Pij taking value1(the feature being selected) after sigmoid transformation and is limited by the maximum velocity parameter V_{max}. A particle is updated in each generation by following their personal best position P_{best} and global best position of the population called G_{best} according to the following two equations:

$$V_{ij} = wV_{ij} + c_1 rand() \cdot (Pbest_{ij} - P_{ij}) + c_2 rand() \cdot (Gbest_{ij} - P_{ij})$$

$$P_{ij} = \begin{cases} 0, & if\, \rho \geq sig(V_{ij}) \\ 1, & if\, \rho < sig(V_{ij}) \end{cases}$$

where c_1 and c_2 are the acceleration coefficients, P_{ij} is the j^{th} element of then dimensional vector Pi. rand() Produces a random number drawn from the normal distribution between 0 and 1. ρ is a random number selected from the uniform distribution in[0,1] as well. The function sig (V_{ij}) is a sigmoid limiting transformation function.

IV. CLASSIFIERS

4.1 k-Nearest Neighbors

The k-nearest neighbors (k-NN) is one of the oldest and simplest non-parametric classification algorithms. Despite its simplicity, it has many advantages and it may give competitive performance compared to many other classification methods. And this algorithm has been successfully applied to a broad range of problems and has numerous variations. This algorithm classifies an unknown sample by comparing it to its k nearest neighbors among a set of known samples, where k is a positive and typically small integer. Firstly, the distances between the unknown sample and all the training samples which mean a set of known samples are calculated. To calculate the distance, various techniques can be used such as Euclidean distance, which is used in the current study, or Mahalanobis distance. After all the distances are calculated, they are sorted and nearest k samples are determined. With a voting scheme among them, i.e. using the majority of the class of nearest neighbors, the class of the unknown sample is assigned.

4.2 Fuzzy K nearest neighbor

K nearest neighbor (KNN) classifies a testing sample according to its K nearest neighbor in the training samples with known classification labels. The sample is then assigned to the class that has the maximum number of neighbors. Fuzzy K nearest neighbor (FKNN) extends the traditional KNN by introducing a fuzzy membership function and distance weight. Fuzzy membership can be used to estimate the confidence level for each class and the weight gives the distance to k nearest neighbors a certain power for the testing sample. The membership value ui(x) to class i is calculated by the following formula:

$$u_i(x) = \frac{\sum_{j-1}^{k} u_i(x^{(j)})\left(\|x - x^{(j)}\|^{-2/(m-1)}\right)}{\sum_{j-1}^{k}\left(\|x - x^{(j)}\|^{-2/(m-1)}\right)} \qquad i = 1,\ldots,c$$

where k is the number of neighbors used and m is the fuzzifier variable which determines how the membership varies with distance.

V. CONCLUSIONS

Biomarker identification is one of the major research area in medical domain. It is a challenging task because of the high dimension (thousand of genes) and low amount of samples. A few biomarker genes related to each type of cancer are identified from the frequency analysis in multiple runs. Firest we have to use dimensionality reduction to transform a large dataset into a reduced set of features. By combining filter and wrapper approaches, we can improve the computation efficiency and robust against overfitting. These methods can be applied to any feature selection tasks in other research fields.

VI. REFERENCES

[1] T.Abeel, T.Helleputte, Y.VandePeer, P.Dupont, Y.Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26(3)(2010)392–398.

[2] P.Maji, S.K.Pal, Fuzzy rough sets for information measures and selection of relevant genes from microarray data, IEEETrans. Syst.ManCybern.PartB – Cybern.40(3)(2010)741–752.

[3] I. Guyon, A.Elisseeff, An introduction to variable and feature selection, J.Mach. Learn. Res.3 (2003)1157–1182.

[4] Hall, M.A., "Correlation-based feature selection machine learning", Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[5] Li, J. and Wong, L., Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns, Bioinformatics, 18:725{734, 2002.

[6] Li, J. and Wong, L., Emerging patterns and geneexpression data, Genome Informatics, 12:3{13,2001.

[7] Dietterich, T. [1998] An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization, Machine Learning 1-22.

[8] P. Fortina, S. Surrey, and L. J. Kricka, "Molecular diagnostics: hurdles for clinical implementation", Trends Molecular Medicine, vol. 8, pp. 264-266, 2002.

[9] A. Blum and P. Langley, "Selection of relevant features and examples in machine learing," Artif. Intell.,vol. 97, no. 1/2, pp. 245–271, 1997.

[10] S. Bandyopadhyay, U. Maulik, and D. Roy, "Gene identification: Classical and computational intelligence approaches," IEEE Trans. Syst., Man, Cybern. C, vol. 38,no. 1, pp. 55–68, Jan. 2008.

[11] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science,
Engineering and Information Technology (IJCSEIT),Vol.2, No.2, April 2012

[12] D.Lavanya, Dr.K.Usha Rani,.," Analysis of feature selection with classification: Breast cancer datasets", Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

[13] Thiemjarus .S, B. P. L. Lo, Laerhoven K.V and G.Z.Yang ,(2005). Feature Selection for Wireless Sensor Networks. In Proc of 1st International Workshop on wearable and Implatable Body Sensors Networks.

[14] A.Su, J.Welsh, L.Sapinoso, S.Kern, P.Dimitrov, H.Lapp, P.Schultz,S.Powell,C.Moskaluk, J.Frierson, G.Hampton, Molecular classification of human carcinomas by use of gene expression signatures,CancerRes.61(20)(2001) 7388–7393.

[15] Ling C., Bolun C., and Yixin C., (2011). Image Feature Selection Based on Ant Colony Optimization.[16] Jin Y., Syed S. R. A., and Paul H. A., (2005). A hybrid feature selection strategy for image defining features: towards interpretation of optic nerve images.

[17] Vasantha M., Dr.V.S. Bharathi, Dhamodharan R, (2010). Medical Image Feature, Extraction, Selection And Classification, International Journal of Engineering Science and Technology. 2(6), 2071-2076.

[18] A. Sharma, S. Imto and S. Miyano, A top-r feature selection algorithm for microarray gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 3 (2012), 754–764.

[19] H.S. Shon, K.S. Yang C.W. Yoo and K.H. Ryu, Feature selection method using WF-LASSO for gene expression data analysis, in: the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, 2011, pp. 522–524.

[20] Q.H. Hu, D.R. Yu and Z.X. Xie, Neighborhood classifiers, Expert Systems with Applications 34 (2008), 866–876.

[21] L. Yu, Y. Han and M.E. Berens, Stable gene selection from microarray data via sample weighting, IEEE/ACM Transactions on Computational Biology and Bioinformatics 9 (2012), 262–272.