A Survey on Different Feature Selection Methods for Microarray Data Analysis

Varuna Tyagi
M.Tech (IT-2nd SEM)
Department of Information Technology
Amity University
Noida, India

Anju Mishra
Assistant Professor
Department of Information Technology
Amity University
Noida, India

ABSTRACT

In the field of medical science diseases diagnosis by Tissue microarrays is one of the active areas of research .There are various gene selection techniques in the literature. Gene selection provides genes subsets that are capable to describe in which category those gene are (active, hyperactive or silent).Various application areas like combinatorial chemistry, text mining, multivariate imaging, or bioinformatics are using huge data sets. The problem has been addressed of selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA micro-arrays for cancer classification. Usually till now survey paper discuss various conventional & evolutionary methods of gene selection like filters, wrappers methods.

Keywords

Features, Genes informative, conventional, evolutionary, SVM.

1. INTRODUCTION

GENE expression microarray (GEM) experiments collect critical biological information collecting biological data from samples like tissues, cell lines. Recorded GEM data hold gene-wise information across all samples in the observation. Recently thousands of genes is measured and recorded simultaneously. In many perspectives these samples can be different under observation. To find the relevant genes for a particular target is an important area of research. These genes are called informative genes. The discovery of informative genes is important for the physician for judge a patients and for the company that are making drugs .in the last few years, a lot of effort has been put in the development of solution for the informative genes discovery. Till now the task is very challenging and some evolutionary approaches is invented to beat the conventional Approaches .The literature of features selection for informative genes discovery is vital. The main goal of this survey is to provide an important family of approaches that is applied in most of the gene selection methods. This survey paper provide a picture of, conventional approaches like filter, wrapper approach [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] [14], [15], [16], [17], [19],[20],[21],[22], [23] are used in section 2. Section 3 provides evolutionary approach [18].

2.CONVENTIONAL APPROACHES

DNA microarray technology provides a facility to monitor thousands of genes simultaneously. If a problem has class than DNA microarray technology can analyse it, these are the problems that have two classes (disease/non-disease, stress/control, Knock-out/wild type) etc. When a conventional approach is used for analysing gene expression data, data mining algorithm is used for detecting genes that are expressed differently. Most of them are as follows:

2.1 Entropy-Based

By this method [1], [2] 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. For a detailed description of the algorithm, please refer to [5, 6, 7] or http://sdmc.lit.org.sg/gedm/Preprocessing.html.

2.2 x^2-Statistics and correlation based feature selection methods

The Chi-Squared (χ^2) method [8] and the Correlation-based Feature Selection (CFS) method [4] are maid on the top of the entropy method. **B**y measuring their chi-squared statistic with respect to the classes, the χ^2 method evaluates features individually. The method first requires its range to be discretized into several intervals using, the entropy-based discretization method for a numeric attribute. By the formula $x^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(Aij-Eij)^2}{Eij}$, The x^2 value of an attribute is defined ,where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the i^{th} interval, j^{th}class, R_i the number of samples in the i^{th} interval,c_j the number of samples in the j^{th} class, N the total number of samples, and E_{ij} the expected frequency of A_{ij} (E_{ij} = R_i *C_j/N) .After getting the x^2 value of all considered features, by putting largest one on first position as proposed these values can be sort, The more important feature is that which have larger x^2value. For 1 degree of freedom at 5% significant level is 3.841 the value of critical x [9].

Another approach of feature selection is CFS method. The worth of subsets of features ranks with this method. Feature subsets are huge, Best first search is used by CFS. Good features subsets hold highly correlated features. CFS construct a matrix then calculate score of subset by the formula: $Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{K+k(K-1)\overline{r_{ff}}}}$, where $Merit_S$ is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ is the average feature-class correlation, and $\overline{r_{ff}}$ is the average feature inter-correlation.

2.3 T-Statistics and MIT Correlation

Based on t-statistics it is widely used feature selection technique. With a data set S consisting of m expression vectors: $x^i = (x_1^i, \dots \dots \dots, x_n^i)$, where 1 ≤ i ≤ m, m is the number of samples, and n is the number of features measured. Each sample is labelled with Y Є {+1,-1} (for classes, such as T-ALL vs. OTHERS1). For each feature x_j, the mean μ_j^+ (resp.μ_j^-) and the standard deviation δ_j^+ $(resp.\ \delta_j^-)$ using only the samples labelled +1 (resp. -1) are calculated. Then a score T(x_j) can be obtained by T(x_j) $= \frac{|\mu_j^+ - \mu_j^-|}{\sqrt{\frac{(\delta_j^+)^2}{n_+} + \frac{(\delta_j^-)^2}{n_-}}}$ where n_+ (resp. n-) is the number of samples labelled as +1 of (resp. -1).For making selection as proposed feature with highest score is taken.. The score is defined as [3]: MIT(x_j) = $\frac{|\mu_j^+ - \mu_j^-|}{\delta_j^+ + \delta_j^-}$.

2.4 Filter methods–a ranking approach

As proposed by Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe most filter methods consider the problem of Feature Selection as a ranking problem. The solution is that genes that have the higher score rest are discarded. Scenario is described below:

- To quantify the differences in expression between rank features and different groups of samples in descending order of the estimated score, scoring function s(x) is used.
- The statistical significance is estimated like confidence intervals of the estimated scores.
- The highest ranked features is selected which are statistically called the most informative genes.
- To ensure that the selected subset of informative genes is valid or not.

Scoring Functions—Assigning Relevance Indices to Features - Scoring functions represent the core of ranking methods and they are used to assign a relevance index to each feature[11].

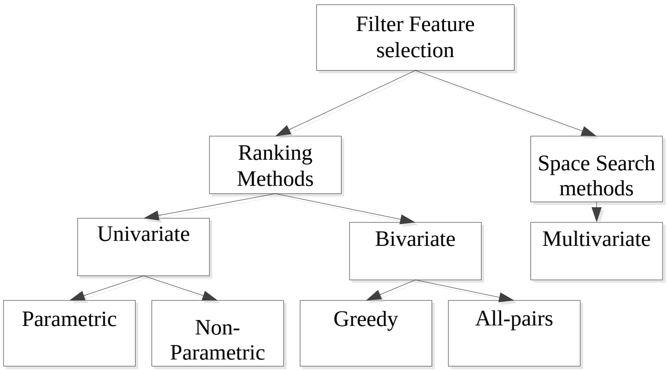

Fig 1: Proposed taxonomy for filter Feature Selection methods.

2.4.1 Ranking methods for feature selection

As proposed in Fig1 it is divided in the following approaches.

2.4.1.2 Univariate methods

According to [13], univariate methods, there are two approaches of feature selections as:

(a) Parametric methods

The data are drawn from a given probability distribution, on these some more or less explicit assumption these method are based.

(b) Nonparametric methods

In this method data is drawn on the bases of some unknown distribution. To quantify the difference in expression between classes based on some estimate scoring function is used.

2.4.1.3 Bivariate ranking methods

According to their discrimination power between two or more conditions Ranking pairs of genes can be performed either using a “greedy strategy” or “all pair strategy.”

(a) Greedy strategies

First rank all genes by individual ranking (using one of the criteria provide by uni-variate ranking methods); subsequently the highest scoring gene$g_i$is paired with the gene g_j that gives the highest gene pair score. After selecting first pair, g_s that is next highest ranked gene paired with the gene g_r ,it maximizes the pair score, and so on. In [14], a greedy gene pair ranking method has been proposed where t-test was performed on first rank gene individually, when the pair score is measured how well the combination of pairs is distinguished between two populations.

(b) All pairs strategies

In all pairs strategies unlike the greedy methods, by computing the pair score for all pair , all pair’s strategies examine all possible gene pairs.

2.4.2 Filter methods–space search approach

It is an optimization strategy. That will provide most informative and least redundant subset of features among the whole set. This strategy follow three main steps described as follows:

1. To optimize Define a cost function.
2. To find the subgroup of features, which optimizes the cost function Use an optimization algorithm.
3. To ensure selected subsets of genes is valid or not.

A filter method [15] provides gene selection independently, For using the data for classification analysis first pre-process the Microarray dataset.

2.5 Wrapper method

As proposed by Hong Hu1, Jiuyong Li1, Hua Wang1, and Grant Daggard[15] Contains a gene selection method within a classification algorithm. Wrapper method is not as efficient as filter method because an algorithm runs on original high dimensional array. However, Kohavi and John [16] have discovered that the accuracy of filter method could improve over filter method by wrapper method. By this it proves that accuracy of chosen genes depend upon selected gene. A wrapper method examples is SVMs [17], SVM uses a recursive feature elimination technique under a greedy fashion to destroy the features iteratively until a largest amount of separation is reached

2.5.1 Clustering techniques

AS proposed by Daxin Jiang Chun Tang Aidong Zhang[19] ,Proximity measurement is a method that measure the similarity (or distance) between two data objects. After taking gene expression data's genes as an object, can be formalized as numerical vectors $\vec{o_i} = \{a_{ij} | 1 \leq j \leq p\}$, where o_{ij} is the value of the j^{th} feature for the i^{th} data object and p is the number of features. The proximity between two objects o_i and o_j is measured by a *proximity function* of corresponding vectors $\vec{o_i}$ and$\vec{o_j}$.

By Euclidean distance the distance between two data object can be measure as:

$$\text{Euclidean}\ (o_{i,} o_j)\ = \sqrt{\sum_{d=1}^{p} \left(o_{id} - o_{jd}\right)^2}$$

The overall shapes of gene expression patterns have a greater interest than the individual magnitudes of each feature. For shifting or scaled patterns Euclidean distance does not score well [20]. For this problem, standardization of each object vector is done with zero mean and variance one before calculating the distance [21, 22, 23].

To measures the similarity between the shapes of two expression patterns an alternate measure is *Pearson's correlation coefficient*, given two data objects o_i and o_j, Pearson's correlation coefficient is defined as

$$\text{Pearson}\ \ (o_{i,} o_j) = \frac{\sum_{d=1}^{p} (o_{id} - \mu_{oi})(o_{jd} - \mu_{o_j})}{\sqrt{\sum_{d=1}^{p} (o_{id} - \mu_{oi})^2}\ \sqrt{\sum_{d=1}^{p} (0_{jd} - \mu_{oj})^2}}$$

Where μ_{oi} and μ_{oj} are the means $\vec{o_i}$ and $\vec{o_j}$, respectively. Pearson's correlation coefficient shows each object as a random variable with diagnosis and measures the similarity between two objects after calculating the linear relationship between the distributions of the two corresponding random variables.

3. EVOLUTIONARY METHODS

As proposed by R. Debnath, and T. Kurita [18] Evolution by natural selection is called an evolutionary approach That genes will be previously selected in the competition to their competitors whose phenotypic effect promote their own propagation successfully. For the benefit of genes this process produces adoptions that promote the reproductive success of the organism that contains the same gene (kin altruism and green-beard effects), or the detriment to the other genes of the genome its own propagation (intra-genomic conflict).

3.1 Evolutionary algorithm

As proposed by R. Debnath, and T. Kurita [18], The optimization techniques and the stochastic search that have been developed over the last 30 years called Evolutionary algorithm. The evolutionary algorithm in general form is shown below:

```
1. Generate initial population, evaluate fitness
 2. While stop condition not satisfied do
      3. Produced next population by
             4. Selection
             5. Recombination
        6. Evaluate fitness
7. End while
```

As proposed, the evolutionary algorithm, whose effectiveness can be determined by using them as features in an SVM classifier maintains a population of predictors. In the population the initial predictors are randomly constructed. The proposed method selects and recombines new features based on leave-one-out error bounds on SVM such as radius margin bound Instead of applying crossover and mutation operations, frequency of occurrence, Jaakkola-Haussler bound and Opper-Winther bound of the features in the evolutionary approach. As proposed in a predictor the number of features is parameter that shall be explore experimentally in the following section. By choosing optimum parameters of SVMs high performance of evolutionary SVM is obtained. Where the evolutionary SVM is applied the k-fold cross validation is used as an estimator of the generalization ability on a k-fold cross validation set and on several different k-fold cross validation sets then the generalization ability of the selected feature is tested. Using both the maximum number of generations and the criteria of no improvement of maximum fitness value of the population the termination criteria is defined. The predictor that contains the best subset of genes for the classification task will be that contain the highest fitness. The term is used as follows:

3.1.1 Error Bound Effect

In every generation, in equations the right hand side of any equation is calculated to observe the effect on error bound of each gene in each predictor. Let us denote T_m is the bound value of m genes on a predictor and T_{m-1}^{4} is the bound value of all genes except gene i. Then, T_{m-1}^{4} for all i are calculated. The $T_{m-1}^{j} < T_{m-1}^{k}$ means removing gene j from the predictor can reduce error Bound much than removing gene k. Thus genes j will small T_{m-1}^{j} should be deleted in the next generation.

3.1.2 Gene Frequency

Let us denote Z_i^j be the frequency of occurrence of selected gene I at generation j. Initially all Z_i^0 is set to 0.At any generation j, if gene I is selected then

$$z_i^j = z_i^{j-1}+1$$

This frequency is calculated for each predictor separately.

3.1.3 Gene Deletion

As proposed remove those genes which can reduce the error bound much and in the previous generations which are selected a few. It calculate the scoring function as

$$T_I = \epsilon T_{m-1}^i + (1- \epsilon)^{Z_i^j}/_l$$

Where ϵ Є [0, 1] is a trade-off between bound value and the frequency of occurrence in the previous generations. Gene i with the minimum T_i will be deleted from the predictor.

3.1.4 Fitness Function

For evaluating the system The fitness function is the only guide. For designing evolutionary SVM There are two objectives. One is to maximize the classification accuracy C_a of the k-fold cross-validation and the other is to minimize the number N_f of selected genes. If S represents the set of parameters to be evolved in the whole system, the fitness function is defined as follows:
max t(S) = $(1 - w_f)C_a$(S) $- w_f N_f$ (S)
where $w_f \in$ [0, 1] is a control parameter between classification accuracy and the number of selected genes.

3.1.5 Proposed Algorithm

The proposed by R. Debnath, and T. Kurita [18] algorithm is described below:

1) A population E0 of n predictors $\{G_1, G_2, ..., G_n\}$ is created. A predictor G_i is a subset of m features (genes) $\{g_1, g_2, ..., g_m\}$ initially created randomly. Evaluate the fitness values of all predictors.
2) UNTIL termination criteria NOT satisfied DO:
3) For each predictor $G_i \in E_k$, create a new predictor $G_i^{'}$.
    1) Delete p genes from G_i as described in Subsection 3.1.3.
    2) Add p genes chosen randomly to keep the size of the feature set the same, i.e.,size(Gi) = size(G'_i). Compute the frequency of the selected genes as described in Subsection 3.1.4.
    3) Compute fitness function for the new predictor $G_i^{'}$.
4) Create a new population E_{k+1} by replacing all new $G_i^{'}$.
5) Replace some worse predictors of the new population E_{k+1} based on classification accuracy by some best predictors from the previous generation.

For replacing worse predictors , to create new $G_i^{'}$ like cross-fold validation technique merge the features of the selected best predictors from the previous generation and then randomly select features from the merge-feature set. The best hyper parameters for each predictor will be obtained and For a set of SVM hyper parameters this procedure will be performed. From this procedure we will get n feature sets. From the n sets we will choose N_{best} top-rank features in terms of occurrence frequency.

4. CONCLUSION

This paper presents a study on different existing gene selection methods for gene expression Microarray data for classification by classifier. Gene selection methods has two approaches first approach is conventional approach which contain Entropy-Based, the χ^2-Statistics and Correlation-Based Feature Selection Methods, T-Statistics and MIT Correlation, a filter method, Wrapper Method, Clustering Techniques. The second approach is evolutionary approach under which evolutionary algorithm comes. In this paper we discussed a brief description of most of the feature selection methods. That shows feature selection is a challenging task.

5. REFERENCES

[1] Huiqing Liu, Jinyan Li, Limsoon Wong, A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns ,Laboratories for Information Technology, 21 Heng Mui Keng Terr, 119613 Singapore Genome Informatics 13: 51{60 (2002).

[2] Fayyad, U. and Irani, K., Multi-interval discretization of continuous-valued attributes for classification learning, Proc. 13th International Joint Conference on Arti_cial Intelligence, 1022{1029,1993

[3] Golub, T. R. et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, 286, 531{537, 1999.

[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. et al., Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukaemia (ALL) patients, Bioinformatics, in press.

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

[8] Liu, H. and Setiono, R., Chi2: Feature selection and discretization of numeric attributes, Proc.IEEE 7th International Conference on Tools with Arti_cial Intelligence, 338{391, 1995.

[9] Sandy R, Statistics for Business and Economics, McGrawHill, 1989.

[10] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter,Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe, A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis, ieee/acm transactions on computational biology and bioinformatics, vol. 9, no. 4, july/august 2012.

[11] S. H. Cha, Comprehensive Survey on Distance/Similarity Measures Between Probability

Density Functions, Int'l J. Math. Models and Methods in Applied Sciences, vol. 1, no. 4, pp. 300-307, 2007.

[12] J. Cohen, "The Earth is Round (p < .05)," Am. Psychologist, vol. 38, pp. 997-1003, 1994.

[13] Y. Saeys, I. Inza, and P. Larran˜ aga, A Review of Feature Selection Techniques in Bioinformatics, Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[14] T. Bø and I. Jonassen, New Feature Subset Selection Procedures for Classification of Expression Profiles, Genome Biology, vol. 4, no. 4, pp. research0017.1-research0017.11, 2002.

[15] Hong Hu1, Jiuyong Li1, Hua Wang1, and Grant Daggard2, Combined Gene Selection Methods for Microarray Data Analysis.

[16] R. Kohavi and G. H. John, Wrappers for feature subset selection Artificial Intelligence,97(1-2):273–324, 1997.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik., Gene selection for cancer classification using support vector machines Machine Learning, 46(1-3):389–422, 2002.

[18] R. Debnath, and T. Kurita, An Evolutionary Gene Selection Method for Microarray Data Based on SVM Error Bound Theories, Neuroscience Research Institute AIST, Tsukuba, Ibaraki, 305-8568, Japan

[19] Daxin Jiang Chun Tang Aidong Zhang, Cluster Analysis for Gene Expression Data: A Survey, Department of Computer Science and Engineering State University of New York at Buffal.

[20] Wang, Haixun, Wang, Wei, Yang, Jiong and Yu, Philip S,. Clustering by Pattern Similarity in Large Data Sets, In SIGMOD 2002, Proceedings ACM SIGMOD International Conference on Management of Data, pages 394–405, 2002.

[21] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture. Nature Genet, pages 281–285, 1999.

[22] Smet, Frank De, Mathys, Janick, Marchal, Kathleen, Thijs, Gert, Moor, Bart De and Moreau, Yves. Adaptive quality-based clustering of gene expression profiles Bioinformatics, 18:735–746, 2002.

[23] Shamir R. and Sharan R., Click: A clustering algorithm for gene expression analysis. In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00). AAAIPress., 2000.