A study on Feature Selection Techniques in Bio-Informatics

S.Nirmala Devi
Department of Master of Computer Applications
Guru Nanak College
Chennai, India
csnirmala77@yahoo.co.in

Dr. S.P Rajagopalan
Department of Master of Computer Applications
Dr.M.G.R Educational and Research Institute
Chennai, India
sasirekaraj@yahoo.co.in

***Abstract*— The availability of massive amounts of experimental data based on genome-wide studies has given impetus in recent years to a large effort in developing mathematical, statistical and computational techniques to infer biological models from data. In many bioinformatics problems the number of features is significantly larger than the number of samples (high feature to sample ratio datasets) and feature selection techniques have become an apparent need in many bioinformatics applications.**
This article provides the reader aware of the possibilities of feature selection, providing a basic taxonomy of feature selection techniques, discussing its uses, common and upcoming bioinformatics applications.

Keywords- ***Bio-Informatics; Feature Selection; Text Mining; Literature Mining; Wrapper; Filter Embedded Methods.***

I. INTRODUCTION

During the last ten years, the desire and determination for applying feature selection techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for model building. The high dimensional nature of the modeling tasks in bioinformatics, going from sequence analysis over microarray analysis to spectral analyses and literature mining has given rise to a wealth of feature selection techniques are presented in the field.

The application of feature selection techniques is focused in this article. While comparing with other dimensionality reduction techniques like projection and compression, feature selection techniques do not alter the original representation of the variables, but merely select a subset of the representation. Thus, it preserves the original semantics of the variables and Feature selection is also known as variable selection, feature reduction, attribute selection or variable subset selection.

Feature selection helps to acquire better understanding about the data by telling which the important features are and how they are related with each other and it can be applied to both supervised and unsupervised learning. The interesting topic of feature selection for unsupervised learning (clustering) is a more complex issue, and research into this field is recently getting more attention in several communities and the problem of supervised leaning is focused here, where the class labels are known already.

The main aim of this study is to make aware of the necessity and benefits of applying feature selection techniques. It provides an overview of the different feature selection techniques for classification by reviewing the most important application fields in the bioinformatics domain, and the efforts done by the bioinformatics community in developing procedures is highlighted. Finally, this study point to some useful data mining and bioinformatics software packages that can be used for feature selection.

II. FEATURE SELECTION TECHNIQUES

Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy (inclusion of irrelevant features can introduce noise into the data, thus obscuring relevant features). It is worth noting that even though some machine learning algorithms perform some degree of feature selection themselves (such as classification trees); feature space reduction can be useful even for these algorithms. Reducing the dimensionality of the data reduces the size of the hypothesis space and thus results in faster execution time.

As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications. The objectives of feature selection are

(a) to avoid over fitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering
(b) to provide faster and more cost-effective models
(c) to gain a deeper insight into the underlying processes that generated the data.

Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset [1], as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset.

There are three types of feature subset selection approaches: depending on how they combine the feature selection search with the construction of the classification model: filters, wrappers and embedded methods which perform the features selection process as an integral part of a machine learning (ML) algorithm. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.

A. *Filter Methods*

These methods do not require the use of a classifier to select the best subset of features. They use general characteristics of the data to evaluate features. Filter techniques use the intrinsic properties of the data to assess the relevance of features. In many cases the low-scoring features are removed and feature relevance score is calculated, then this subset is given as input to the classification algorithm.

They are pre-processing methods. They attempt to assess the merits of features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by ranking them through compression techniques or by computing correlation with the output.

Advantages of filter techniques are that they are independent of the classification algorithm, computationally simple and fast and easily scale to very high-dimensional datasets. Feature selection needs to be performed only once, and then different classifiers can be evaluated.

Disadvantages of filter methods is that they ignore the interaction with the classifier i.e., the search in the feature subset space is separated from the search in the hypothesis space. Each feature is considered separately and compared to other types of feature selection techniques it lead to worse classification performance thereby ignoring feature dependencies. A number of multivariate filter techniques were introduced in order to overcome the problem of ignoring feature dependencies.

B. *Wrapper methods*

These methods assess subsets of variables according to their usefulness to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. The problem boils down to a problem of stochastic state space search. Examples are the stepwise methods proposed in linear regression analysis. This method embeds the model hypothesis search within the feature subset search. A search procedure of possible feature subsets is defined and various subsets of features are generated and evaluated. The training and testing a specific classification model evaluation produces a specific subset of features. A search algorithm is then 'wrapped' around the classification model to search the space of all feature subsets. These search methods can be divided in two classes deterministic and randomized search algorithms.

Advantages of Wrapper Method include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies.
Disadvantages are that they have a higher risk of over fitting than filter techniques.

III. Applications in Bioinformatics

A. *Feature Selection for Sequence Analysis*

A multistage process that includes the determination of a sequence (protein, carbohydrate, etc.), its fragmentation and analysis, and the interpretation of the resulting sequence information. This information is useful in that it: (a) reveals the similarities of homologous genes, thereby providing insight into the possible regulation and functions of these genes; and (b) leads to a better understanding of disease states related to genetic variation. New sequencing methodologies, fully automated instrumentation, and improvements in sequencing-related computational resources contribute to the potential for genome-size sequencing projects.

In the context of feature selection, two types of problems can be distinguished: signal and content analysis. Signal analysis focuses on identifying the important motifs in the sequence, such as gene regulatory elements or structural elements. On the other hand content analysis focuses on the broad characteristics of a sequence, such as tendency to code for proteins or fulfillment of a certain biological function and feature selection techniques are then applied to focus on the subset of relevant variables.

1) *Content Analysis*

In early days of bioinformatics the prediction of subsequence's that code for proteins has been focused. Many versions of Markov models were developed because many features are extracted from a sequence, and most dependencies occur between adjacent positions. Interpolated Markov model was introduced to deal with limited amount of samples [2], and the high amount of possible features. This method used filter method to select only relevant features and interpolation between different orders of the Markov model to deal with small sample sizes. Later Interpolated Markov Model was extended to deal with non-adjacent feature dependencies, resulting in the interpolated context model (ICM), which crosses a Bayesian decision tree with a filter method ($\lambda 2$) to assess feature relevance. Recognition of promoter regions and the prediction [3], of microRNA targets are the use of FS techniques in the domain of sequence analysis.

2) *Signal Analysis*

For the recognition of short, more or less conserved signals in the sequence many sequence analysis methods are used and also to represent the binding sites for various proteins or protein complexes. Regression Approach is the common approach to find regulatory motifs and to relate motifs to gene expression levels to search for the motifs that maximize the fit to the regression model [4], Feature selection is used .In 2003 to find discriminative motifs a classification approach is chosen . This method uses the threshold number of misclassification (TNoM) to score genes for relevance to tissue classification. From the TNoM score, to represents the significance of each motif a P-value is calculated and according to their P-value Motifs are then sorted.

Another line of research is performed in the context of the gene prediction setting, where structural elements such as the translation initiation site (TIS) and splice sites are modeled as specific classification problems. In future research, FS techniques can be expected to be useful for a number of challenging prediction tasks, such as identifying relevant features related to alternative TIS and alternative splice sites .

B. Feature Selection for Microarray Analysis

The human genome contains approximately 20,000 genes. At any given moment, each of our cells has some combination of these genes turned on, and others are turned off. Scientists can answer this question for any cell sample or tissue by gene expression profiling, using a technique called microarray analysis. Microarray analysis involves breaking open a cell, isolating its genetic contents, identifying all the genes that are turned on in that particular cell and generating a list of those genes.

During the last decade, the introduction of microarray datasets stimulated a new line of research in bioinformatics. Microarray data pose a great challenge for computational techniques, because of their small sample sizes and their large dimensionality. Furthermore, additional experimental complications like noise and variability render the analysis of microarray data an exciting domain. A dimension reduction technique was realized in order to deal with these particular characteristics of microarray data and soon their application became a de facto standard in the field. Whereas in 2001, the field of microarray analysis was still claimed to be in its infancy a considerable and valuable effort has since been done to contribute new and adapt known FS methodologies.

1) The Univariate Filter Paradigm

This Method is simple yet efficient because of the high dimensionality of most microarray analyses, fast and efficient FS techniques such as univariate filter methods have attracted most attention. The prevalence of these techniques has dominated the field and now comparative evaluations of different FS techniques and classification over DNA microarray datasets focused on the univariate .This domination of the this approach can be explained by a number of reasons:

(a) The univariate feature rankings output is intuitive and easy to understand;
(b) the objectives and expectations that bio-domain experts have when wanting to subsequently validate the result by laboratory techniques or in order to explore literature searches is fulfilled by the output of the gene ranking . The experts could not feel the need for selection techniques that take into account gene interactions;
(c) multivariate gene selection techniques the needs extra computation time .
(d) the possible unawareness of subgroups of gene expression domain experts about the existence of data analysis techniques to select genes in a multivariate way;

The detection of the threshold point in each gene that reduces the number of training sample misclassification and setting a threshold on the observed fold-change differences in gene expression between the states under study are some of the simplest heuristic rule for the identification of differentially expressed genes. A wide range of new univariate feature ranking techniques has since then been developed. These techniques can be divided into two classes: parametric and model-free methods.

Parametric methods assume a given distribution from which the observations (samples) have been generated. t-test and ANOVA are the two samples among the most widely used techniques in microarray studies, although the usage of their basic form, possibly without justification of their main assumptions, is not advisable [5]. To deal with the small sample size and inherent noise of gene expression datasets include a number of t- or t-test like statistics (differing primarily in the way the variance is estimated) and a number of Bayesian frameworks are the modifications of the standard t-test. Regression modeling approaches and Gamma distribution models are the other types of parametrical approaches found in the literature.

Due to the uncertainty about the true underlying distribution of many gene expression scenarios, and the difficulties to validate distributional assumptions because of small sample sizes, non-parametric or model-free methods have been widely proposed as an attractive alternative to make less stringent distributional assumptions. The Wilcox on rank-sum test [6], between-within classes sum of squares (BSS/WSS) [7], and the rank products method [8]. Are the model-free metrics of statistics field have demonstrated their usefulness in many gene expression studies.

These model-free methods uses random permutations of the data to estimate the reference distribution of the statistics allowing the computation of a model-free version of the associated parametric tests. These techniques deal with the specificities of DNA microarray data, and do not depend on strong parametric assumptions. Their permutation principle partly alleviates the problem of small sample sizes in microarray studies and enhancing the robustness against outliers.

2) The multivariate paradigm for filter, wrapper and embedded techniques

Univariate selection methods have certain restrictions and it leads to less accurate classifiers by, e.g. not taking into account gene–gene interactions. Thus, researchers have proposed techniques that try to capture these correlations between genes. Correlation-based feature selection (CFS) [9], and several variants of the Markov blanket filter method are the application of multivariate filter methods ranges from simple bivariate interactions towards more advanced solutions exploring higher order interactions. The two other solid multivariate filter procedures are Minimum Redundancy-Maximum Relevance (MRMR) [10], and Uncorrelated Shrunken Centroid (USC) [11], algorithms highlighting the advantage of using multivariate methods over univariate procedures in the gene expression domain.

Feature selection uses an alternative way to perform a multivariate gene subset selection, incorporating the classifier's bias into the search and thus offering an opportunity to construct more accurate classifiers. The scoring function is another characteristic of any wrapper procedure and is used to evaluate each gene subset found. As the 0–1 accuracy measure allows for comparison with previous works, the vast majority of papers use this measure. However, recent proposals advocate the use of methods for the approximation of the area under the ROC curve [12], or the optimization of the LASSO (Least Absolute Shrinkage and Selection Operator) model [13].For screening different types of errors in many biomedical scenarios ROC curves certainly provide an interesting evaluation measure.

The embedded capacity of several classifiers to discard input features and thus propose a subset of discriminative genes has been exploited by several authors. A random forest (a classifier that combines many single decision trees) is an example to calculate the importance of each gene. The weights of each feature in linear classifiers, such as SVMs and logistic regression are used by embedded FS techniques and these weights are used to reflect the relevance of each gene in a multivariate way, and thus allow for the removal of genes with very small weights.

Due to the lesser degree embedded approaches and higher computational complexity of wrapper, these techniques have not received as much interest as filter proposals. However univariate filter method is an advisable practice to pre-reduce the search space, and only then apply wrapper or embedded methods, hence fitting the computation time to the available resources.

C. *Mass Spectra Analysis*

For disease diagnosis and protein-based biomarker profiling the emerging new and attractive framework is the Mass spectrometry technology (MS). A mass spectrum sample is characterized by thousands of different mass/charge (m/ z) ratios on the x-axis, each with their corresponding signal intensity value on the y-axis. A typical MALDI-TOF low-resolution proteomic profile can contain up to 15,500 data points in the spectrum between 500 and 20, 000 m/z, and the number of points even grows using higher resolution instruments.

For data mining and bioinformatics purposes, it can initially be assumed that each m/ z ratio represents a distinct variable whose value is the intensity. The data analysis step is severely constrained by both high-dimensional input spaces and their inherent sparseness, just as it is the case with gene expression datasets. Although the amount of publications on mass spectrometry based data mining is not comparable to the level of maturity reached in the microarray analysis domain, an interesting collection of methods has been presented in the last 4–5 years.

The following crucial steps is to extract the variables that will constitute the initial pool of candidate discriminative features and starting from the raw data, and after an initial step to reduce noise and normalize the spectra from different samples . Some studies employ the simplest approach of considering every measured value as a predictive feature, thus applying FS techniques over initial huge pools of about 15,000 variables, up to around 1,00,000 variables. The elaborated peak detection and alignment techniques are the great deal of current studies performs aggressive feature extraction procedures. These procedures tend to seed the dimensionality from which supervised FS techniques will start their work in less than 500 variables. To set the computational costs of many FS techniques to a feasible size the feature extraction step is thus advisable in these MS scenarios. Univariate filter techniques seem to be the most common techniques used which is Similar to the domain of microarray analysis, even though the use of embedded techniques is certainly emerging as an alternative. The other parametric measures such as notable variety of non-parametric scores and F-Test have also been used in several MS studies. Although the t-test maintains a high level of popularity. Multivariate filter techniques on the other hand, are still somewhat underrepresented.

In MS studies Wrapper approaches have demonstrated their usefulness by a group of influential works. in the major part of these papers different types of population-based randomized heuristics are used as search engines: genetic algorithms [14], particle swarm optimization (Ressom et al., 2005) and ant colony procedures [15].To discard input features an increasing number of papers uses the embedded capacity of several classifiers. Variations of the popular method originally proposed for gene expression domains using the weights of the variables in the SVM-formulation to discard features with small weights, have been broadly and successfully applied in the MS domain .Based on a similar framework, to rank the features by the weights of the input masses in a neural network classifier. The alternative embedded FS strategy is the embedded capacity of random forests and other types of decision tree-based algorithms.

IV. Dealing With Small Sample Domains

Small sample sizes and their over fitting and inherent risk contain a great challenge for many modeling problems in bioinformatics. Two initiatives have emerged in the context of feature selection (i.e.) the use of adequate evaluation criteria, and the use of stable and robust feature selection models in response to this novel experimental situation.

A. *Adequate evaluation criteria*

Several papers have warned about the substantial number of applications not performing an independent and honest validation of the reported accuracy percentages. In such cases, a discriminative subset of features is often selected by the users using the whole dataset. This subset is used to estimate the accuracy of the final classification model thus testing the discrimination rule on samples that were already used to propose the final subset of features. The need for an external feature selection process in training the classification rule at each stage of the accuracy estimation procedure is gaining space in the bioinformatics community practices. Furthermore, novel predictive accuracy estimation methods with promising characteristics, such as bolstered error estimation have emerged to deal with the specificities of small sample domains.

B. *Ensemble feature selection approaches*

An ensemble system, on the other hand is composed of a set of multiple classifiers and performs classification be selecting from the predictions made by each of the classifiers. Since wide research has shown that ensemble systems are often more accurate than any of the individual classifiers of the system alone and it is only natural that ensemble systems and feature selection would be combined at some point.

Instead of choosing one particular FS method different FS methods can be combined using ensemble FS approaches and accepting its outcome as the final subset Based on the evidence that there is often not a single universally optimal feature selection technique and due to the possible existence of more than one subset of features that discriminates the data equally well [11], model combination approaches such as boosting have been adapted to improve the robustness and stability of final, discriminative methods [16]. To assess the relevance of each feature in an ensemble FS the methods based on a collection of decision trees (e.g. random forests) can be used. Although the use of ensemble approaches requires additional computational resources, we would like to point out that they offer an advisable framework to deal with small sample domains, provided the extra computational resources are affordable.

V. Feature selection in Upcoming Domains

A. *Single nucleotide polymorphism analysis*

A **single-nucleotide polymorphism** (**SNP**, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between members of a species or paired chromosomes in an individual. Single nucleotide polymorphisms (SNPs) are mutations at a single nucleotide position that occurred during evolution and were passed on through heredity, accounting for most of the genetic variation among different individuals. SNPs are number being estimated at about 7 million in the human genome and it is the forefront of many disease-gene association studies. The important step towards disease-gene association is selecting a subset of SNPs that is sufficiently informative but still small enough to reduce the genotyping overhead. Typically, the number of SNPs considered is not higher than tens of thousands with sample sizes of about 100.

In the past few years several computational methods for htSNP selection (haplotype SNPs; a set of SNPs located on one chromosome) have been proposed. One approach is based on the hypothesis that the human genome can be viewed as a set of discrete blocks that only share a very small set of common haplotypes. The aim of this approach is to identify a subset of SNPs that can either explain a certain percentage of haplotypes or atleast distinguish all the common haplotypes. Another common htSNP selection approach is based on pairwise associations of SNPs, and tries to select a set of htSNPs such that each of the SNPs on a haplotype is highly associated with one of the htSNPs [17]. The remaining SNPs can be reconstructed and it is the third approach considering htSNPs as a subset of all SNPs. The idea is to select htSNPs based on how well they predict the remaining set of the unselected SNPs.

B. *Text and literature mining*

It is the emerging as a promising area for data mining in biology. Text mining or text data mining, or text analytics, refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. Bag-of-Words (BOW) representation is one important representation of text and documents where the variable represents each word in the text representation of the text may lead to very high dimensional datasets, pointing out the need for feature selection techniques.

In the field of text classification the application of feature selection techniques is common and the application in the biomedical domain is still in its infancy. A large number of feature selection techniques that were already developed in the text mining community for tasks such as biomedical document clustering and classification and it will be of practical use for researchers in biomedical literature mining .

VI. Fs Software Packages

Table I shows an overview of existing software In order to provide the interested reader with some pointers to existing software packages implementing a variety of feature selection methods. The software is organized into four sections: general purpose FS techniques, techniques tailored to the domain of microarray analysis, techniques specific to the domain of mass spectra analysis and techniques to handle SNP selection and all software packages mentioned are free for academic use. For each software package, the main reference, implementation language and website is shown.

For each software package, the main reference, implementation language and website is shown.

TABLE I SOFTWARE FOR FEATURE SELECTION

General Purpose FS software			
WEKA	Java	Witten and Frank(2005)	http://www.cs.waikato.ac.nz/ml/weka
Fast Correlation Based Filter	Java	Yu and Liu(2004)	http://www.public.asu.edu/~huanliu/FCBF/FCBFsoftware.html
MLC++	C++	Kohavi et al.(1996)	http://www.sgi.com/tech/mlc
Feature selection Book	Ansi C	Liu and Motoda(1998)	http://public.asu.edu/~huanliu/FSbook
Microarray analysis FS software			
SAM	R.Excel	Tusher et al.(2001)	http://www-stat.stanford.edu/~tibs/SAM/
PCP	C,C++	Buturovic(2005)	http://pcp.sourceforge.net
GALGO	R	Trevino & Falciani(2006)	http://www.bip.bham.ac.uk/bioinf/galgo.html
GA-KNN	C	Li et al(2001)	http://dir/niehs.nih.gov/microarray/datamining/
Nudge(Bioconductor)	R	Dean & Raftery(2005)	http://www.bioconductor.org/
Qvalue(Bioconductor)	R	Storey(2002)	http://www.bioconductor.org/
DEDS(Bioconductor)	R	Yang et.al(2005)	http://www.bioconductor.org/
Mass Spectra analysis FS software			
GA-KNN	C	Li et al(2004)	http://dir.niehs.nih.gov/microarray/datamining/
R-SVM	R,C,C++	Zhang et al.(2006)	**http://www.hsph.harvard.edu/bioinfocore/RSVMhome/R-SVM.html**
SNP analysis FS software			
CHOISS	C++, Perl	Lee and Kang(2004)	http://biochem.kaist.ac.kr/choiss.htm
WCLUSTAG	Java	Sham et al.(2007)	http://bioinfo.hku.hk/wclustag

VII. CONCLUSIONS AND FUTURE PERSPECTIVES

In this article, it is reviewed the main contributions of feature selection research in a set of well-known bioinformatics applications. Te large input dimensionality and the small sample sizes are the two main issues emerge as common problems in the bioinformatics domain. Researchers designed FS techniques to deal with these problems in bioinformatics, machine learning and data mining.

During the last years a large and fruitful effort has been performed in the adaptation and proposal of univariate filter FS techniques. In general, it is observed that many researchers in the field still think that filter FS approaches are only restricted to univariate approaches. The proposal of multivariate selection algorithms can be considered as one of the most promising future lines of work for the bioinformatics community.

A second line of future research is The development of especially fitted ensemble FS approaches to enhance the robustness of the finally selected feature subsets is the second line of future research. In order to alleviate the actual small sample sizes of the majority of bioinformatics applications, the further development of such techniques, combined with appropriate evaluation criteria, constitutes an interesting direction for future FS research.

SNPs, text and literature mining, and the combination of heterogeneous data sources are the other interesting opportunities for future FS research will be the extension towards upcoming bioinformatics domains. While in these domains, the FS component is not yet as central as, e.g. in gene expression or MS areas, I believe that its application will become essential in dealing with the high-dimensional character of these applications.

ACKNOWLEDGMENT

I would like to thank the anonymous reviewers for their constructive comments, which significantly improved the quality of this review.

REFERENCES

[1] Daelemans,W.,et al.(2003) Combined optimization of feature seelection and algorithm parameter interaction in machine learning of language. In Proceedings of the 14th European Conference on Machine Learning (ECML – 2003), pp. 84-95.

[2] Salzberg., et.al(1998) Microbial gene identification using interpolated markov models. Nucleic Acids Res., 26, 544–548.

[3] Saeys,Y., et al. (2007) In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi, and protists. Bioinformatics, 23, 414–420.

[4] Keles,S., et al. (2002) Identification of regulatory elements using a feature selection method. Bioinformatics, 18, 1167–1175.

[5] Jafari,P. and Azuaje,F. (2006) An assessment of recently published gene

expression data analyses: reporting experimental design and statistical factors,BMC Med. Inform. Decis. Mak., 6, 27.

[6] Thomas,J., et al. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles.Genome Res., 11, 1227–1236.

[7] Dudoit,S., et al. (2002) Comparison of discriminant methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97, 77–87.

[8] Breitling,R., et al. (2004) Rank products: a simple, yet powerful, new method todetect differentially regulated genes in replicated microarray experiments.FEBS Lett., 573, 83–92.

[9] Wang,Y., et al. (2006) Tumor classification based on DNA copy number aberrations determined using SNPS arrays. Oncol. Rep., 5, 1057–1059.

[10] Ding,C. and Peng,H. (2003) Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the IEEE Conference on Computational Systems Bioinformatics, pp. 523–528

[11] Yeung,K. and Bumgarner,R. (2003) Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biol., 4, R83.

[12] Ma,S. and Huang,J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. Bioinformatics, 21, 4356–4362.

[13] Ghosh,D. and Chinnaiyan,M. (2005) Classification and selection of biomarkers in genomic data using LASSO. J. Biomed. Biotechnol., 2005,147–154.

[14] Li,T., et al. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression . Bioinformatics, 20, 2429–2437.

[15] Ressom,H., et al. (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics, 23, 619–626.

[16] Ben-Dor,A., et al. (2000) Tissue classification with gene expression profiles. J. Comput. Biol., 7, 559–584

[17] Carlson,C., et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet., 74, 106–120..

[18] Margaret H.Dunham S.Sridhar (2008) Data Mining Introductory and Advanced Topics.

[19] Efron,B., et al. (2001) Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc., 96, 1151–1160.

[20] Kohavi,R., et al. (1996) Data mining using MLC++: a machine learning library in C++. In Tools with Artificial Intelligence, IEEE Computer Society Press, Washington, DC, pp. 234–245.

[21] Inza,I., et al. (2000) Feature subset selection by Bayesian networks based optimization. Artif. Intell., 123, 157–184.

[22] Sofie Van Landeghem (2008), Extracting Protein –Protein Interactions from Text using Rich Feature Vectors and Feature Selection. 77-84.

[23] Michael Gutkin (2009) , A method for feature selection in gene expression-based disease classification.

[24] Varshavsky,R., et al. (2006) Novel unsupervised feature filtering of biological data. Bioinformatics, 22, e507–e513.

[25] Jake Y. Chen , Stefano Lonardi (2009), Biological Data Mining.

[26] Pawel Smialowski, *Bioinformatics (2010),* Pitfalls of supervised feature selection,oxford journals *26(3): 440-443.*