Perspective of Feature Selection Techniques in Bioinformatics

Satish Kumar David[1], Mohammad Khalid Siddiqui[2]
[1]IT Department, [2]Department of Basic Sciences
[1,2]Strategic Center for Diabetes Research, King Saud University, Riyadh, Saudi Arabia

Abstract

The availability of massive amounts of experimental data based on genome-wide association and mass spectroscopy studies have given motivation 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 data sets) and feature selection techniques have become an apparent need in many bioinformatics applications. In addition to the large pool of techniques that have already been developed in the data mining fields, specific applications in bioinformatics have led to a wealth of newly proposed techniques. This assessment provides the aware of the possibilities of feature selection, providing a basic taxonomy of feature selection techniques, discussing their use, variety and potential in a number of both common as well as upcoming bioinformatics applications.

Keywords

Bioinformatics; Feature Selection; Text Mining; Wrapper; Genotype analysis.

Introduction

Now a day's interest for using Feature Selection (FS) techniques in bioinformatics becoming compulsion for model building from being just example. The modeling tasks going to spectral analysis and text mining from sequence analysis over microarray analysis in bioinformatics. FS helps to acquire better understanding about the data's important features and their relationship type and can be applied to supervised (classification, prediction) and unsupervised (Clustering) learning [1]. The original representation of the variables does not vary in FS techniques but dimensionality reduction techniques such as projection and compression can vary the original representation of the variables. From an informatics perspective, the process of selecting differentially expressed genes is readily achieved via data-mining techniques known as Feature Selection. It is an important step in the data-mining process aims to find representative optimal feature subsets that meet desired criteria. The key consideration in this review is FS techniques application and the idea is to bring awareness of the requirements and benefits of using FS techniques. This article also will give an idea about few useful data mining and bioinformatics software packages used for FS.

In microarray data analysis, one criterion for a desired feature subset would be a set of genes whose expression patterns vary significantly when compared across different sample groups. The resulting subset can then be used to further analysis such as building a diagnostic classifier. Problem of selecting some subset of a learning algorithm's input variables upon which it should focus attention, while ignoring the rest (Dimensionality Reduction). Several pattern recognition techniques alone do not handle with large amounts of irrelevant features. Pattern recognition techniques and FS techniques jointly work effectively in many applications [2].A large number of features enhances the model's flexibility, but makes it prone to over fitting. The FS objectives are

(a) To increase the speed of learning algorithm's
(b) To improve the accuracy of classifier on new data
(c) To remove redundant features from dataset.

In classification context approaches for FS techniques tasks are: filters, wrappers and embedded methods [3].

Filter Approach

FS is based on an evaluation criterion for quantifying how well feature (subsets) discriminate the two classes in Filter techniques. Filters assess the relevance of features. Relevance score calculated and low scored are removed and then this subset is input to classification algorithm. Only once FS needs to be performed and then different classifiers evaluated [4]. Improved scalability, simple and fast is the advantages of filter techniques. Disadvantages of filter techniques are classifiers performance may be non-optimal features [5]. To prevail over the problem of overlooking feature dependencies, numbers of multivariate filter techniques were introduced.

Wrapper Approach

Wrapper techniques are iterative approach, many feature subsets are scored based on classification performance. Running a model on the subset wrappers use a search algorithm to search through the space of possible features and evaluate each subset. Wrappers have higher over fitting risk and can be computationally expensive. These search methods assess subsets of variables according to their usefulness to a given classifier [6]. Based on search method the wrapper methods divided into two kinds a)randomize [7,8] b)Greedy [9]. Advantages of Wrapper techniques are improving the performance of given

classifier. Disadvantages of Wrapper techniques are computationally intensive, high cost and poor scalability.

Embedded Approach

Embedded techniques are specific to a model. These methods use all the variables to generate and analyze the model to recognize the importance of the variables [10]. FS is part of classifier's training procedure (e.g. decision trees). Consequently, they directly link variable importance to the learner used to model the relationship. Attempt to jointly or simultaneously train both a classifier and a feature subset. Often optimize an objective function that jointly rewards accuracy of classification and penalizes use of more features. Advantages of Embedded technique are less computationally intensive. Disadvantage of embedded technique is classifier dependent classifier.

Literature Mining

Automated methods for knowledge retrieval from the text are known as literature mining. Most knowledge is stored in terms of texts, both in industry and in academia. In biology promising area for data mining is literature mining [11]. Word based system Bag-of-Words (BOW) representation is changing set of words linearly structured into unstructured which may lead to very high dimensional datasets and the need for feature selection techniques [12]. BOW based models use statistical weights based on term frequency, document frequency, passage frequency, and term density. BOW disregards grammatical structure, layout free representation and context dependent. Literature mining developed for document clustering, classification and researcher's practical use.

Sequence Analysis

Sequence analysis is the modern operation in computational biology. This operation find out which part of the biological sequences is alike and which part differs during medical analysis and genome mapping processes. The sequence analysis implies subjecting a DNA to sequence databases, sequence alignment, repeated sequence searches, or other methods in bioinformatics [13]. New sequencing methodologies, fully automated instrumentation, and improvements in sequencing-related computational resources greatly contributed for genome-size sequencing projects. Multistage process contains the purpose of sequence (protein), its fragmentation, analysis and resulting sequence information. This information reveals similarities of homologous genes and its regulation and function of the gene, leads to a better understanding of disease states related to gene variation [14].

Microarray Analysis

Human genome contains approximately 30,000 genes [15]. Each of our cells has some combination of these genes active at any given instant and others inactive. Computation in the microarray data is great challenge because of large dimensionality and small sample size. Multivariate is unsupervised Clustering, Principle component analysis, Classification (statistical learning, discriminant analysis, supervised clustering). According to Jafari considerable and valuable effort has been done to contribute adapt FS, since microarray claims to be infancy [15]. Univariate features ranking techniques has been developed such as parametric and non-parametric (model free). Parametric method assumes given distribution from which samples have been generated. Two samples t-test and ANOVA are mostly used in microarray analysis even though usage not advisable [16]. ANOVA is for measuring the statistical significance of set of independent variables. ANOVA produces the p-value for the features set. ANOVA procedure recommended only for balanced data. Other types of parametric techniques such as regression modelling, Gamma distribution model. Since uncertainty is high in parametric techniques, the model free (non-parametric) techniques proposed. Metrics are from statistical categories(BSS/WSS) [17]. Using random permutation reference distribution of statistics were estimated in model free techniques. Multivariate regressions are Correlation features selection (CFS), minimum redundancy maximum relevance(MRMR). Proposed the use of methods under ROC curve or optimization of LASSO model. ROC gives interesting evaluation measure. Three broad problems in microarray analysis: a) class discovery (unsupervised classification), b) class comparison (differential gene expression), c) class prediction (supervised classification).

Genotype analysis

In the genome wide association study (GWAS) a large number of data have been generated for SNP analysis, its range from 100 to 1000 SNP. These SNP analysis is import to look the relation between phonotypic with genotypic data to relate the different disease condition. Different approaches were used based on data mining and genetic algorithm [18]. A weighted decision tree, a correlation-based heuristic are used for selecting significant genes. The goal of feature selection for SNPs can be achieved with supervised and unsupervised methods such as clustering, neighborhood analysis, applying classification algorithm and eliminating the lowest weight features can pruned DNA gene expression data sets by eliminating insignificant features [19]. The significant gene/SNP set in cross-validation accuracy was increased by 10% over the baseline measurements and the specificity increased by 3.2% over baseline measurements. Block free approach for tagging SNPs. the selection of tagging SNPs can be partitioned into the three following steps:

a. Determining neighborhoods of linkage disequilibrium: Find out which sets of SNPs can be meaningfully used to infer each other.

b. Tagging quality assessment: Define a quality measure that describes how well a set of tagging SNPs captures the variance observed.

c. Optimization: Minimize the number of tagging SNPs.

The disadvantage of block free approach is not always straightforward definition of blocks and no consensus on how blocks must be formed. It is based only on the local correlations. To avoid computational complexity, did not look for subsets of SNPs but discard redundant markers using FS technique. It can give better performance on

large data set using exhaustive search to short chromosomal regions but this does not guarantee optimal solutions. In the genotyping the huge data generated and related between the SNP and LD (Linkage disequilibrium) was used by the block based approach [20].

Mass Spectroscopy Methodology

Mass Spectroscopy analysis is for protein-based biomarker profiling and disease diagnosis. Two different types of mass spectroscopy methodology used for analysis. The common method of sorting ions is the Time-Of-Flight (TOF) analyzer. In TOF analyzer ions are collected to an ion trap, and then accelerated with one push into an empty chamber with an electrical field in it. An instrument MALDI-TOF (Matrix-Assisted Laser Desorption and Ionization Time-Of-Flight) low resolution can contain upto 15,500 data points in spectrum and number of points can even grows for higher resolution instruments. MALDI-TOF is the most popular techniques presently employed for detecting quantitative or qualitative changes of proteins [21]. Mass spectrometry measures two properties of ion mixtures in the gas phase under a vacuum environment: the mass/charge ratio (m/z) of ionized proteins in the mixture and the number of ions present at different m/z values. Thus the mass spectrometry for a sample is a function of the molecules and used to test for presence or absence of one or more. The general pipeline is show in Figure 1, which includes three steps.

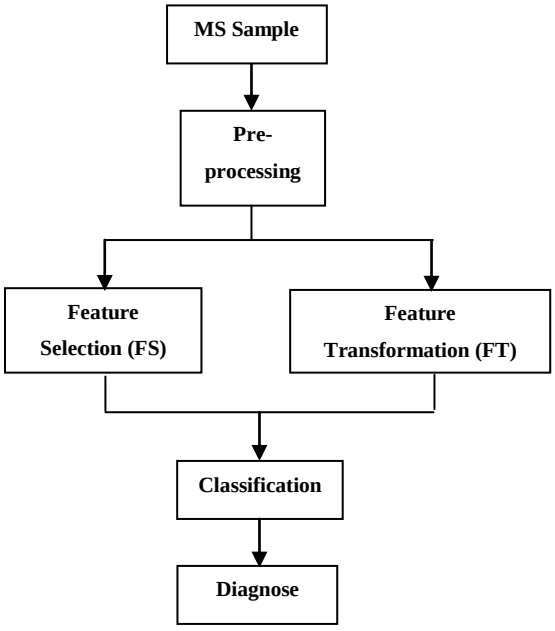

Figure 1. Pipeline of pattern analysis of MS data

Firstly, the MS data is pre-processed. Then, two kinds of dimension reduction methods are accepted. One is called feature transformation (FT). FT methods construct new features as functions that express relationships between the initial features. The other kind of methods is called feature selection (FS). The FS methods output a subset of the original input features without transforming them, such as t-test, sequential forward selection (SFS), boosting approaches, etc. The last step is classification, which gives the results of diagnosis, such as SVM, KNN, decision tree, etc.

As Somorjai et al. explained the data analysis steps is constrained by both high dimension input spaces and their inherent sparseness [20]. Several studies employ the simplest approach of considering every measured value as predictive features, so applying FS technique over 15000 variables upto around 100000 variables [22].

Ensemble feature selection

An ensemble system is composed of a set of multiple classifiers and performs classification be selecting from the predictions made by each of the classifiers. Ensemble FS derived from decision tree and used to assess relevance of each features. 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. Composed of set of multiple classifiers and performs classification be selecting from predictions made by each of the classifiers. Frequently a single FS technique is not optimal and redundant subset of feature data [23]. Therefore, Ensemble FS have been incorporated to improve the methods strength and methods stability [24]. Additional computational resources are required to use ensemble FS and if additional resources are affordable, ensemble FS offer framework to deal with small sample.

Conclusion and Future perspective

In this review we assess feature selection techniques in bioinformatics applications. Table1 shows software's packages, their main reference and website shown. These software packages are free for academic use. We found issues and problems of small sample size and large dimensionality in data mining. Feature Selection techniques designed to deal with these problems. Productive effort has been performed in the proposal of univariate filter FS techniques. Future research is the development of ensemble Feature Selection approaches to enhance the robustness of selected feature subset and literature mining. Interesting opportunities towards genotype analysis is needed.

References

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

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

[3] Y. Saeys et al. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics/btm344, Vol. 23 no. 19, pages 2507-2517, June 2007.

[4] Yu,L. and Liu,H. (2004) Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res., 5, 1205–1224.

[5] Ben-Bassat,M. (1982) Pattern recognition and reduction of dimensionality. In Krishnaiah,P. and Kanal,L., (eds.) Handbook of Statistics II, Vol. 1. North-Holland, Amsterdam. pp. 773–791.

[6] Inza,I., et al. (2000) Feature subset selection by Bayesian networks based optimization. Artif.Intell., 123, 157–184.
[7] X. Wang, J. Yang, X. Teng, W. Xia, J. Richard, Feature selection based on rough sets and particle swarm optimization, Pattern Recognition Letters 28 (2007) 459–471.
[8] M. Ronen, Z. Jacob, Using simulated annealing to optimize feature selection problem in marketing applications, European Journal of Operational Research 171 (2006) 842–858.
[9] S.F. Cotter, K. Kreutz-Delgado, B.D. Rao, Backward sequential elimination for sparse vector selection, Signal Processing 81 (2001) 1849–1864.
[10] Eugene Tuv et al (2009) Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination.Journal of Machine Learning Research 10, Pages 1341-1366, July 2009.
[11] M. Grobelnik et al."Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining," 2000.
[12] Tellex, S., B. Katz, J. Lin, A. Fernandes, and G. Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 41–47.
[13] Pawel Smialowski et al., *Data and text mining* Pitfalls of supervised feature selection, Vol. *26 no. 3 2010, pages 440–443 doi:10.1093/bioinformatics/btp621*
[14] Chikina MD, Troyanskaya OG (2011) Accurate Quantification of Functional Analogy among Close Homologs. PLoS Comput Biol 7(2): e1001074. doi:10.1371/journal.pcbi.1001074
[15] Somorjai,R., et al. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics, 19, 1484–1491.
[16] 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.
[17] Sima,C., et al. (2005) Superior feature-set ranking for small samples using bolstered error estimation. Bioinformatics, 21, 1046–1054.
[18] Shital C. Shah, Andrew Kusiak, Data mining and genetic algorithm based gene/SNP selection, Artificial Intelligence in Medicine (2004) 31, 183—196
[19] Raychaudhuri S, Sutphin PD, Chang JT, Altman RB. Basicmicroarray analysis: grouping and feature reduction. Trends Biotechnol 2001;19(5):189—93.
[20]Tu Minh Phuong, Zhen Lin Russ B. Altman, Choosing SNPs Using Feature Selection, Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)
[21] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198-207, 2003
[22] Li,L. et al. (2004) Applications of the GA/KNN method to SELDI proteomics data. Bioinformatics, 20, 1638–1640.
[23] Yeung,K. and Bumgarner,R. (2003) Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biol., 4, R83.
[24] Ben-Dor,A., et al. (2000) Tissue classification with gene expression profiles. J. Comput. Biol., 7, 559–584
[25]Alyssa J Porter et al. (2009), ProMerge - A ToolKit for Data Capture and Integration in Differential Proteomics, 3rd Annual Conference In Quantitative Genomics, November 11-13, 2009, Joseph B Martin Conference Center, Boston MA, US
[26] Li, L., Weinberg, C.R., Darden, T.A. and Pedersen, L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131-1142.
[27] Aharoni A. and Vorst O. 2002. DNA microarrays for functional plant genomics. Plant Mol. Biol. 48(1):99-118.
[28] Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. 2003 Feb;34(2):374-8.
[29] Bradley Efron and Tibshirani, 2007, On testing the signi_cance of sets of genes, Annals of Applied Statistics vol 1.
[30] Trevino &Falciani (2006), "GALGO: an R package for multivariate variable selection using genetic algorithms." Bioinformatics 22(9): 1154-6
[31] Nema Dean (2006), The Normal Uniform Di_erential Gene Expression (nudge) detection package.
[32] Yang et.al (2011), Bioconductor's DEDS package
[33] Rodrigo Alvarez-Gonzalez, et al. (2011), Discriminant Fuzzy Pattern to Filter Di_erentially Expressed Genes.
[34] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
[35] Kohavi et al. (1996), Data Mining with MLC++. A broad view with a large comparison of many algorithms in MLC++ on the large UC Irvine datasets.Received the IEEE Tools with Artificial Intelligence Best Paper Award, 1996.
[36] Lei Yu, Yue Han, and Michael E Berens(2011). "Stable Gene Selection from Microarray Data via Sample Weighting".*IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press, 2011

Table 1.Softwares for Feature Selection

Software Names [References]	Mass Spectra analysis FS Software
ProMerge[25]	http://www.hsph.harvard.edu/research/bioinfocore/resources/software/index.html
GA/KNN[26]	http://www.niehs.nih.gov/research/resources/software/gaknn/index.cfm
Microarray analysis FS Software	
GA/KNN[26]	http://www.niehs.nih.gov/research/resources/software/gaknn/index.cfm
GeneMaths XT [27]	http://www.applied-maths.com/download/software.htm
TM4[28]	http://www.tm4.org/
SAM[29]	http://www-stat.stanford.edu/~tibs/SAM/
GALGO[30]	http://biptemp.bham.ac.uk/vivo/galgo/AppNotesPaper.htm
Nudge[31]	http://www.bioconductor.org/packages/release/bioc/html/nudge.html
DEDS[32]	http://www.bioconductor.org/packages/release/bioc/html/DEDS.html
DFP[33]	http://www.bioconductor.org/packages/release/bioc/html/DFP.html
General Purpose FS Software	
WEKA[34]	http://www.cs.waikato.ac.nz/ml/weka/
MLC++ [35]	http://www.sgi.com/tech/mlc/utils.html
FCBF[36]	http://www.public.asu.edu/~huanliu/FCBF/FCBFsoftware.html
Genomic Analysis	
SLAM[13]	http://bio.math.berkeley.edu/slam/
Multiz[13]	http://www.bx.psu.edu/miller_lab/