Feature selection techniques in bioinformatics

Conference Paper · July 2017

CITATIONS
0

READS
89

3 authors, including:

Naser Nematbakhsh
Shahid Ashrafi Esfahani University, Esfahan, I…
83 PUBLICATIONS **212** CITATIONS
SEE PROFILE

Motahareh Nadimi
University of Isfahan
7 PUBLICATIONS **2** CITATIONS
SEE PROFILE

دارای مجوز ISC وزارت علوم
96170 - 10906
کنفرانس ملی پژوهش های نوین در
برق، کامپیوتر و مهندسی پزشکی
۲۷ تیر ماه ۱۳۹۶
دانشگاه آزاد اسلامی
واحد کازرون

Feature selection techniques in bioinformatics

Mohamad Reza Hosseini
Department of Computer Engineering, Shahid Ashrafi Esfahani University, Isfahan, Iran
Soheil.hosseini20@gmail.com

Naser Nematbakhsh
Department of Computer Engineering, Shahid Ashrafi Esfahani University, Isfahan, Iran
n.neemat@yahoo.com

Motahareh Nadimi
Department of Biology, Faculty of Science, University of Isfahan, Isfahan, Iran
M_nadimi68@yahoo.com

Abstract
Machine learning methods are often used to classify objects described by hundreds of attributes; however, as the dimensionality of the data rises, the amount of data required to provide a reliable analysis grows exponentially. A popular approach to this problem of high-dimensional datasets is to search for a projection of the data onto a smaller number of variables (or features) which preserves the information as much as possible. Feature selection is an important step in data mining and is used in various subjects including genetics, medicine, and bioinformatics. 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 and discussing its uses in bioinformatics applications including sequence analysis, microarray analysis, discovering Statistically-Equivalent Feature Subsets in the R Package MXM, classification of pre-miRNAs and Mass spectra analysis.

Keywords: Bioinformatics; Feature Selection; Wrapper; Filter; Embedded Methods.

Introduction

As the world grows in complexity, overwhelming us with the data it generates, data mining becomes the only hope for elucidating the patterns that underlie it (Witten, Frank et al. 2011). The manual process of data analysis becomes tedious as size of data grows and the number of dimensions' increases, so the process of data analysis needs to be computerized (Beniwal and Arora 2012).

The term Knowledge Discovery from data refers to the automated process of knowledge discovery from databases. Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Data mining can also be explained as the non-trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern and so on (Shan, Wei et al. 2009). The knowledge extracted from data mining, allows the user to find interesting patterns and regularities deeply buried in the data to help in the process of decision making (Beniwal and Arora 2012).

Advances in genetics, chemistry, and information technology have allowed researchers to discover biomarkers that are related to the diagnosis of a disease or its treatment (Atkinson, Colburn et al. 2001). DNA contains many biomarkers, but only a small subset of these biomarkers will be related to any specific disease. Therefore, one of the goals of research into these biomarkers is to identify which are relevant to the problem at hand. One of the most common ways of achieving this goal is through ordering or ranking the genes by importance (Awada, Khoshgoftaar et al. 2012).

Ordering the genes by importance is very similar to feature selection (a data preprocessing step from the domain of data mining). A feature is an individual measurable property of the process being observed. By providing a dataset to a feature selection technique, the feature selection algorithm returns the features that are the most important to the problem at hand. In the case of biological and genetics experiments, the problem being studied can be anything from distinguishing between healthy and diseased tissue (Dudoit, Fridlyand et al. 2002) to identifying and accurately classifying between different types of cancer or subtypes of the same cancer (Ben-Dor, Bruhn et al. 2000) to patient response prediction to a drug treatment and many more (Mulligan, Mitsiades et al. 2007).

During the last decade, the motivation for applying feature selection techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for model building. In particular, the high dimensional nature of many modelling 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 being presented in the field (Saeys, Inza et al. 2007, Lazar, Taminau et al. 2012). For examples in gene microarray analysis, the standardized gene expression data can contain hundreds of variables of which many of them could be highly correlated with other variables. The dependent variables provide no extra information about the classes and thus serve as noise for the predictor. This means that the total information content can be obtained from fewer unique features which contain maximum discrimination information about the classes. Hence by eliminating the dependent variables, the amount of data can be reduced which can lead to improvement in the classification performance. In some applications, variables which have no correlation to the classes serve as pure noise might introduce bias in the predictor and reduce the classification performance. This can happen when there is a lack of information about the process being studied. By applying feature selection techniques, we can gain some insight into the process and can improve the computation requirement and prediction accuracy (Chandrashekar and Sahin 2014).

The main aim of this review is to make practitioners aware of the benefits, and in some cases even the necessity of applying feature selection techniques. In this paper we will focus on feature selection methods using supervised learning algorithms and a very brief introduction to feature selection

methods using unsupervised learning will be presented. Therefore, we provide an overview of the different feature selection techniques for classification: we illustrate them by reviewing the most important application fields in the bioinformatics domain, highlighting the efforts done by the bioinformatics community in developing novel and adapted procedures.

Feature Selection Techniques

Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore, feature selection techniques need to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid overfitting and improve model performance and to provide faster and more cost-effective models (Beniwal and Arora 2012).

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; 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 (Devi and Rajagopalan 2011).

The four key steps of a Feature selection process are feature subset generation, subset evaluation, stopping criterion and result validation. The feature subset generation is a heuristic search process which results in the selection of a candidate subset for evaluation. It uses searching strategies like complete, sequential and random search to generate subsets of features (Dunne, Cunningham et al. 2002, K.Sutha and Tamilselvi 2015). The goodness of the generated subset is evaluated using an evaluation criterion. If the newly generated subset is better than the previous subset, it replaces the previous subset with the best subset. These two processes are repeated until the stopping criterion is reached. The final best feature subset is then validated by prior knowledge or using different tests. Fig.1 illustrates the feature selection process (K.Sutha and Tamilselvi 2015).

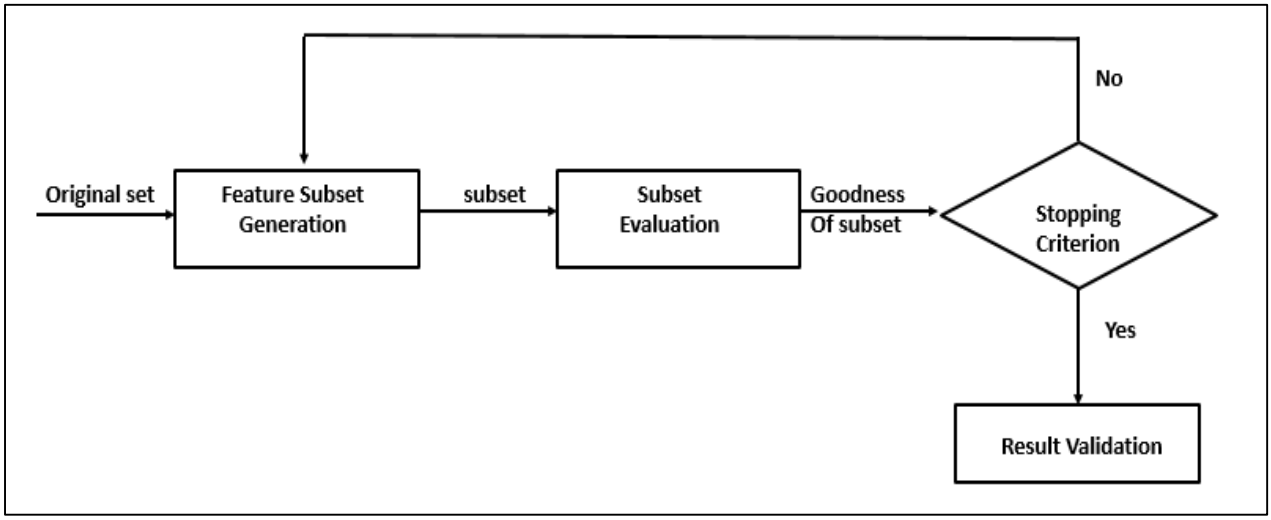


Figure 1. Feature Selection Process (K.Sutha and Tamilselvi 2015)

Feature selection algorithms are separated into three categories (Mwangi, Tian et al. 2014, Tohka, Moradi et al. 2016):

(i) The filters which extract features from the data without any learning involved.

(ii) The wrappers that use learning techniques to evaluate which features are useful.

(iii) The embedded techniques which combine the feature selection step and the classifier construction (Hira and Gillies 2015).

Filter-based feature selection techniques seek to select the optimum feature subset through the use of statistical metrics. There are two main classes of filter-based feature selection: feature ranking and subset evaluation. Feature ranking (Figure 2) (Awada, Khoshgoftaar et al. 2012), sometimes known as univariate feature selection, take each feature separately and tests it for its ability to distinguish between the classes. Subset evaluation looks at the possible subsets of features and tests them for their ability to differentiate between the classes. It should be noted that in order to test each possible subset, one would have to perform $2^n - 1$ tests where n is the number of features per instance. Although search algorithms can reduce this space significantly, they do not resolve the problem of necessitating many evaluations, and introduce the problem of finding local optima which are not the globally-preferred solutions (Saeys, Inza et al. 2007).

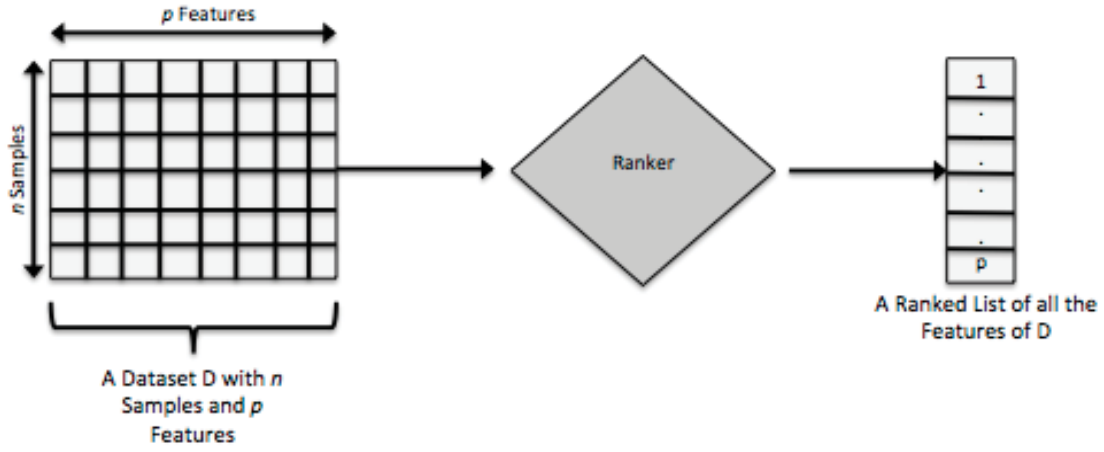


Figure 2. Feature Ranking (Awada, Khoshgoftaar et al. 2012)

Advantages of filter techniques are that they easily scale to very high-dimensional datasets, they are computationally simple and fast, and they are independent of the classification algorithm. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated (Saeys, Inza et al. 2007).

A common disadvantage of filter methods is that they ignore the interaction with the classifier (the search in the feature subset space is separated from the search in the hypothesis space), and that most proposed techniques are univariate. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques. In order to overcome the problem of ignoring feature dependencies, a number of multivariate filter techniques were introduced (Devi and Rajagopalan 2011).

wrapper-based feature selection techniques rely on building classification models to determine the importance of features. Typically, wrapper-based feature selection is performed on subsets (much like filter based subset evaluation), although it may be applied as a ranker. Wrapper methods differ from filter-based feature selection in that they use a learner when evaluating the features, either separately or as subsets. Unfortunately, the building of a learner takes time and would have to be repeated for every test (Hall and Smith 1999). This aspect of wrapper-based feature selection can make the technique very computationally expensive, especially for subset evaluation. For example, if one

wanted to choose the best pair of features from a dataset of 15,154 features, it would require evaluating 114,814,281 classifiers. If it takes 0.1 seconds to develop each classifier it will take 132.9 days of continuous computation to evaluate them all Again, different search techniques can improve this, but introduce their own problems (Awada, Khoshgoftaar et al. 2012).

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 (Devi and Rajagopalan 2011).

Embedded feature selection techniques are found within the learners themselves. This means that when one runs a learner with embedded feature selection, the learner performs feature selection prior to analysis (Awada, Khoshgoftaar et al. 2012). Embedded FS algorithms solve the learning and variable selection problems jointly by optimizing a suitably regularized objective function consisting of a data term and a regularization term whose trade-off is controlled by regularization parameters (Tohka, Moradi et al. 2016). A well-known embedded technique is random forests. A random forest is a collection of classifiers. New random forests are created iteratively by discarding a small fraction of genes that have the lowest importance (D´ıaz-Uriarte and Andr´es 2006). The forest with the smallest amount of features and the lowest error is selected to be the feature subset. A method called block diagonal linear discriminant analysis (BDLDA) assumes that only a small number of genes are associated with a disease and therefore only a small number are needed in order for the classification to be accurate. To limit the number of features it imposes a block diagonal structure on the covariance matrix. In addition, SVMs can be used for both feature selection and classification. Features that do not contribute to classification are eliminated in each round until no further improvement in the classification can be achieved (Hira and Gillies 2015). Support vector machines-recursive feature elimination (SVM-RFE) starts with all the features and gradually excludes the ones that do not identify separating samples in different classes. A feature is considered useful based on its weight resulting from training SVMs with the current set of features. In order to increase the likelihood that only the “best” features are selected, feature elimination progresses gradually and includes cross-validation steps. A major advantage of SVM-RFE is that it can select high-quality feature subsets for a particular classifier. It is however computationally expensive since it goes through all features one by one and it does not take into account any correlation the features might have (Guyon, Weston et al. 2002, Hira and Gillies 2015).

Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods (Saeys, Inza et al. 2007).

So far we have discussed the feature selection techniques and their importance. In the next section we will briefly look at some of feature selection technique`s application in bioinformatics.

Applications in Bioinformatics

A) Feature selection for sequence analysis

Sequence analysis is 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 (Devi and Rajagopalan 2011).

Sequence analysis has a long-standing tradition in bioinformatics. In the context of feature selection, two types of problems can be distinguished: content and signal analysis. Content analysis focuses on the broad characteristics of a sequence, such as tendency to code for proteins or fulfillment of a certain biological function. Signal analysis on the other hand focuses on the identification of important motifs in the sequence, such as gene structural elements or regulatory elements.

Apart from the basic features that just represent the nucleotide or amino acid at each position in a sequence, many other features, such as higher order combinations of these building blocks can be derived, their number growing exponentially with the pattern length k. As many of them will be irrelevant or redundant, feature selection techniques are then applied to focus on the subset of relevant variables (Saeys, Inza et al. 2007).

A-1) Content analysis. In early days of bioinformatics, the prediction of subsequence's that code for proteins has been focused. Because many features can be extracted from a sequence, and most dependencies occur between adjacent positions, many variations of Markov models were developed. Interpolated Markov model was introduced to deal with limited amount of samples, which used interpolation between different orders of the Markov model to deal with small sample sizes, and a filter method (X^2) to select only relevant features (Salzberg 1998). 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 (A. Al-Shahib 2005). The use of feature selection techniques in the domain of sequence analysis is also emerging in a number of more recent applications, such as the recognition of promoter regions and the prediction of microRNA targets (Conilione and Wang 2005, S. Kim 2006).

A-2) Signal Analysis. Many sequence analysis methodologies involve the recognition of short, more or less conserved signals in the sequence, representing mainly 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 (S. Keles 2002). Feature selection can then be used to search for the motifs that maximize the fit to the regression model. In Sinha (Sinha 2003), a classification approach is chosen to find discriminative motifs. The method is inspired by Ben-Dor et al. (2000) who use 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 (H. Liu 2004, Saeys, Inza et al. 2007).

B) Feature selection for microarray analysis

The human genome contains approximately 25,000 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 (Devi and Rajagopalan 2011).

Microarray databases are a large source of genetic data, which, upon proper analysis, could enhance our understanding of biology and medicine. Many microarray experiments have been designed to investigate the genetic mechanisms of cancer, and analytical approaches have been applied in order to classify different types of cancer or distinguish between cancerous and noncancerous tissue. In the last ten years, machine learning techniques have been investigated in microarray data analysis. Several approaches have been tried in order to (i) distinguish between cancerous and noncancerous samples, (ii) classify different types of cancer, and (iii) identify subtypes of cancer that may progress aggressively. All these investigations are seeking to generate biologically meaningful interpretations of complex datasets that are sufficiently interesting to drive follow-up experimentation (Hira and Gillies 2015).

During the last decade, the advent of microarray datasets stimulated a new line of research in bioinformatics. Microarray data pose a great challenge for computational techniques, because of their large dimensionality (up to several tens of thousands of genes) and their small sample sizes (R. Somorjai 2003). 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 feature selection methodologies (Jafari and Azuaje 2006, Saeys, Inza et al. 2007).

Many different feature selection and feature extraction methods exist and they are being widely used. All these methods aim to remove redundant and irrelevant features so that classification of new instances will be more accurate. The most commonly used methods on microarray data analysis are shown in Table 1 (Hira and Gillies 2015).

C) Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets

It is often the case that multiple feature subsets are approximately equally predictive for a given task. Low statistical power due to an insufficient sample size can simply make it impossible to distinguish the predictive performance of two or more signatures in a statistically meaningful way. More intriguingly, the physical process that generates the data could be possibly characterized by a high level of redundancy: several of its components can have similar or identical behavior/scope. Measurements taken over redundant components would be equivalent to each other, and there would be no particular reason for preferring one over the other for inclusion in a predictive subset. This problem is particularly relevant in biology, where nature uses redundancy for ensuring resilience to shocks or adverse events (Lagani, Athineou et al. 2016).

Discovering multiple and statistically equivalent feature subsets has several advantages. First, knowing that multiple equally-predictive subsets actually exist increases the understanding of the specific problem at hand. In contrast, identifying a single subset of relevant features can lead to ignore factors that may play an important role for understanding the dynamics of the problem under study. On more practical terms, equally-predictive subsets may differ in terms of the cost/effort needed for measuring their respective components. Thus, providing multiple, alternative subsets can have a great impact in contexts where some factors may be technically difficult or excessively expensive to measure (Lagani, Athineou et al. 2016)

Table 1: Feature selection methods applied on microarray data (Hira and Gillies 2015)

Methods	Type	Description

t-test feature selection	Filter	It finds features with a maximal difference of mean value between groups and a minimal variability within each group
Correlation-based feature selection (CFS)	Filter	It finds features that are highly correlated with the class but are uncorrelated with each other
Bayesian networks	Filter	They determine the causal relationships among features and remove the ones that do not have any causal relationship with the class
Information gain (IG)	Filter	It measures how common a feature is in a class compared to all other classes
Genetic algorithms (GA)	Wrapper	They find the smaller set of features for which the optimization criterion (classification accuracy) does not deteriorate
Sequential search	Wrapper	Heuristic base search algorithm that finds the features with the highest criterion value (classification accuracy) by adding one new feature to the set every time
SVM method of recursive feature elimination (RFE)	Embedded	It constructs the SVM classifier and eliminates the features based on their "weight" when constructing the classifier
Random forests	Embedded	They create a number of decision trees using different samples of the original data and use different averaging algorithms to improve accuracy
Least absolute shrinkage and selection operator (LASSO)	Embedded	It constructs a linear model that sets many of the feature coefficients to zero and uses the nonzero ones as the selected features.

Recently, algorithms that generate multiple, equivalent feature sets have been developed (A and Aliferis 2010, Lagani, Athineou et al. 2016), including the Statistically Equivalent Signatures (SES) method (Tsamardinos, Lagani et al. 2012), which is implemented in the R (Team 2015) MXM package. SES is a constraint based, feature selection algorithm that attempts to identify multiple, equally-predictive signa- tures, where for signatures we indicate minimal-size sets of features with maximal predictive power. SES subsumes and extends previous work on future selection, particularly the maxmin parent children (MMPC) algorithm and related extensions, by implementing a heuristic method for identifying equivalences among predictors (Lagani, Athineou et al. 2016).

Finally, the MXM package is one of the few open-source code providing implementations of constraint-based feature selection algorithms. The MMPC algorithm has been previously implemented in the bnlearn package (Scutari 2010) along with several Bayes Network learning methods, and the TETRAD software (Landsheer 2010) provides implementations of numerous casual discovery oriented constraint-based methods. The MATLAB library Causal Explorer has been the first software offering feature-selection oriented constraint-based methods, but the code is not open-source (Lagani, Athineou et al. 2016).

D) Classification of pre-miRNAs

MicroRNAs (miRNA) are non-coding RNAs about 21~23 nucleotides (nt) in length, which can play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Bartel

2004). It has been shown that miRNAs usually participate in a set of important life processes, including growth processes, hematopoiesis, organ formation, apoptosis, and cell proliferation. Furthermore, they are closely related to many kinds of human diseases, including cancer (Bushati and Cohen 2007). Due to the difficulty of systematically detecting miRNAs from a genome using existing experimental techniques, computational methods play important roles in the identification of new miRNAs (Xuan, Guo et al. 2011).

Precursor miRNAs (pre-miRNAs) of 60~70 nt have stem-loop hairpin structures, which are an important characteristic feature used in the computational identification of miRNAs. Recently, the ab initio method based on machine learning was presented and applied to distinguish real pre-miRNAs from candidate hairpin sequences (Batuwita and Palade 2009). These classifiers could then classify a candidate sequence as a real pre-miRNA or a pseudo pre-miRNA. However, there are quite a lot extracted features. Not all these features are beneficial, because some features provide little information gain or because some features are redundant relative to other features. Therefore, it is necessary to select the most representative feature subset, which contributes to the improvement of the classification performance (Xuan, Guo et al. 2011).

Genetic algorithm-based feature selection:

Genetic Algorithms (GA) are search algorithms based on natural genetics that provide robust search capabilities in complex spaces, thereby offering a valid approach to problems requiring efficient and effective search processes (Holland 1975). GA is an iterative process that operates on a population, i.e., a set of candidate solutions. Initially, the population is randomly generated. Every individual in the population is assigned, by means of a fitness function, a fitness value that reflects its quality with respect to solving the particular problem (Beniwal and Arora 2012). Furthermore, GA is naturally applicable to feature selection since the problem has an exponential search space.

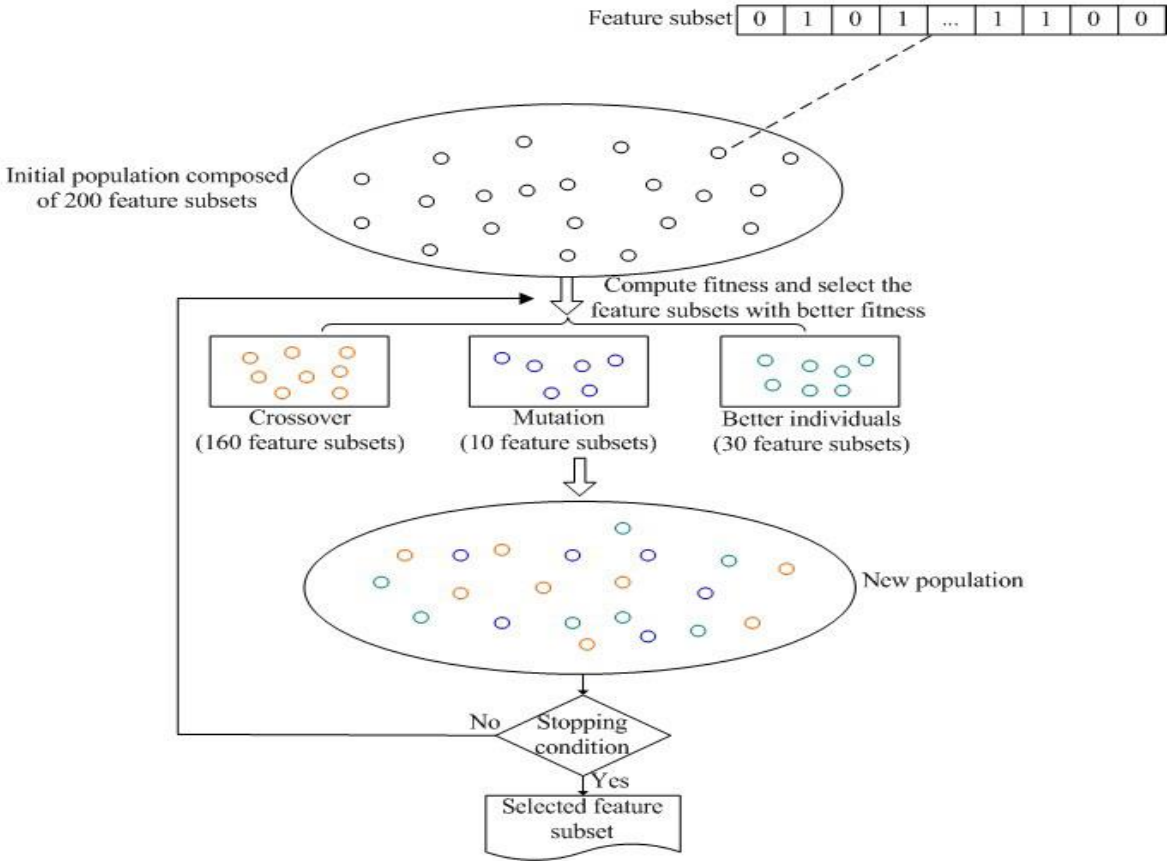
Since the pre-miRNA has 48 dimension features, there are 2^{48} feature subsets. It is not feasible to find the optimal feature subset with an exhaustive method. Therefore, Xuan et al. (2011) propose a feature selection method based on a genetic algorithm. The process of feature selection is shown in Figure 3 (Xuan, Guo et al. 2011).This feature selection algorithm includes conservation statistics, population initialization and genetic iteration. In the conservation calculation step, when the number of real pre-miRNAs is l, the total running time of conservation statistic is $O(l^2)$. In the initialization step, m individuals are constructed. The total running time of initialization is O(m). The genetic iteration consists of calculating the individual fitness, the crossover and the mutation operation. Suppose k is the number of average selected features in all the feature subsets of a population, and the algorithm iterates n times. Thus, the average running time of the iteration step is $O (n*m* k^2)$. The average total running time of feature selection is $O (l^2+m+n*m* k^2) \approx O (n*m* k^2)$.

In addition, in order to construct an SVM classification model, this method just need to select the feature subset only once, and there is no need to select many times. The classification model could then be used to classify real pre-miRNAs and pseudo pre-miRNAs again and again. Therefore, it is worth selecting a representative feature subset to improve classification performance.

E) Mass spectra analysis

For disease diagnosis and protein-based biomarker profiling the emerging new and attractive framework is the Mass spectrometry technology (MS). (Petricoin and Liotta 2003). 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 (Saeys, Inza et al. 2007).

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

Figure 3. Procedure of feature selection based on genetic algorithm (Xuan, Guo et al. 2011)

(R. Somorjai 2003). Although the amount of publications on mass spectrometry based data6 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 (Devi and Rajagopalan 2011).

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 (Devi and Rajagopalan 2011). Some studies employ the simplest approach of considering every measured value as a predictive feature, thus applying feature selection techniques over initial huge pools of about 15 000 variables up to around 100 000 variables (E. Petricoin 2002, Saeys, Inza et al. 2007). 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 feature selection techniques will start their work in less than 500 variables. A feature extraction step is thus advisable to set the computational costs of many FS techniques to a feasible size 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. Although the t-test maintains a high level of popularity (H. Liu 2002) other parametric measures such as F-test and a notable variety of non-parametric scores have also been used in several

MS studies. Multivariate filter techniques on the other hand, are still somewhat underrepresented (Saeys, Inza et al. 2007).

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, particle swarm optimization and ant colony procedures. To discard input features an increasing number of papers uses the embedded capacity of several classifiers (Devi and Rajagopalan 2011).

Conclusions and Future Perspectives

Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Before applying any mining technique, irrelevant attributes needs to be filtered. Filtering is done using different feature selection techniques like wrapper, filter, embedded technique (Beniwal and Arora 2012). The large input dimensionality and the small sample sizes are the two main issues emerge as common problems in the bioinformatics domain. Researchers designed feature selection techniques to deal with these problems in bioinformatics, machine learning and data mining.

In this article, different feature selection methods were described and compared. Their advantages and disadvantages were also discussed. In addition, we presented these methods in a set of well-known bioinformatics applications including sequence analysis, microarray analysis, discovering Statistically-Equivalent Feature Subsets in the R Package MXM, classification of pre-miRNAs and Mass spectra analysis.

Among the existing feature selection algorithms, some algorithms involve only in the selection of relevant features without considering redundancy. Dimensionality increases unnecessarily because of redundant features and it also affects the learning performance. And some algorithms select relevant features without considering the presence of noisy data (K. Sutha and Tamilselvi 2015)

Feature selection techniques show that more information is not always good in machine learning applications. We can apply different algorithms for the data at hand and with baseline classification performance values we can select a final feature selection algorithm. For the application at hand, a feature selection algorithm can be selected based on the following considerations: simplicity, stability, number of reduced features, classification accuracy, storage and computational requirements. Overall applying feature selection will always provide benefits such as providing insight into the data, better classifier model, enhance generalization and identification of irrelevant variables. (Chandrashekar and Sahin 2014).

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.

References:

A, A. S. and C. Aliferis (2010). "Analysis and Computational Dissection of Molecular Signature Multiplicity." PLoS computational biology **6**(5): e1000790.

A. Al-Shahib, e. a. (2005). "Feature selection and the class imbalance problem in predicting protein function from sequence." Appl. Bioinformatics **4**: 195-203.

Atkinson, A. J., et al. (2001). "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework." Clin Pharmacol Ther **69**(3): 89-95.

Awada, W., et al. (2012). "A Review of the Stability of Feature Selection Techniques for Bioinformatics Data." IEEE: 356-363.

Bartel, D. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**: 281-297.

Batuwita, R. and V. Palade (2009). "microPred: effective classification of pre-miRNAs for human miRNA gene prediction." Bioinformatics **25**: 989-995.

Ben-Dor, A., et al. (2000). "Tissue classification with gene expression profiles." Computational Biology **7**(3-4): 559-583.

Beniwal, S. and J. Arora (2012). "Classification and Feature Selection Techniques in Data Mining." International Journal of Engineering Research & Technology (IJERT) **1**(6): 1-6.

Bushati, N. and S. Cohen (2007). " microRNA functions." Annu. Rev. Cell Dev. Biol. **23**: 175-205.

Chandrashekar, G. and F. Sahin (2014). "A survey on feature selection methods." Computers and Electrical Engineering **40**: 16-28.

Conilione, P. and D. Wang (2005). "A comparative study on feature selection for E.coli promoter recognition." Int. J. Inf. Technol. **11**: 54-66.

D´ıaz-Uriarte, R. and S. A. d. Andr´es (2006). "Gene selection and classification of microarray data using random forest." BMC Bioinformatics **7**(3).

Devi, S. N. and S. P. Rajagopalan (2011). "A study on Feature Selection Techniques in Bio-Informatics." International Journal of Advanced Computer Science and Applications **2**(1): 138-144.

Dudoit, S., et al. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data." The American Statistical Association **97**(457): 77-87.

Dunne, K., et al. (2002). "Solution to instability problems with sequential wrapper-based approaches to feature selection." Machine Learning Research.

E. Petricoin, e. a. (2002). "Use of proteomics patterns in serum to identify ovarian cancer." The Lancet **359**: 572-577.

Guyon, I., et al. (2002). "Gene selection for cancer classification using support vector machines." Machine Learning Research **46**(1-3): 389-422.

H. Liu, e. a. (2002). " A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns." Genome Inform. **13**: 51-60.

H. Liu, e. a. (2004). " Using amino acid patterns to accurately predict translation initiation sites. ." In Silico Biol. **4**: 255-269.

Hall, M. A. and L. A. Smith (1999). "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper." in Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference: 235-239.

Hira, Z. M. and D. F. Gillies (2015). "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data." Advances in Bioinformatics: 1-13.

Holland, J. H. (1975). "Adaptation in Natural and Artificial Systems." University of Michigan Press, Ann Arbor.

Jafari, P. and F. Azuaje (2006). "An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors." BMC Med. Inform. Decis. Mak. **6**(27).

K.Sutha and J. J. Tamilselvi (2015). "A Review of Feature Selection Algorithms for Data Mining Techniques." International Journal on Computer Science and Engineering **7**(6): 63-67.

Lagani, V., et al. (2016). "Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets." Statistical Software **10**(2): 1-25.

Landsheer, J. (2010). "The Specification of Causal Models with Tetrad IV: A Review." Structural Equation Modeling: A Multidisciplinary Journal **17**: 703-711.

Lazar, C., et al. (2012). "A survey on filter techniques for feature selection in gene expression microarray analysis." IEEE/ACM Trans Comput Biol Bioinform **9**.

Mulligan, G., et al. (2007). "Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib." Blood: 3177-3188.

Mwangi, B., et al. (2014). "A review of feature reduction techniques in neuroimaging." Neuroinformatics **12**(2): 229-244.

Petricoin, E. and L. Liotta (2003). "Mass spectometry-based diagnostic: the upcoming revolution in disease detection." Clin. Chem. **49**: 533-534.

R. Somorjai, e. a. (2003). "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions." Bioinformatics **19**: 1484-1491.

S. Keles, e. a. (2002). "Identification of regulatory elements using a feature selection method." Bioinformatics **18**: 1167-1175.

S. Kim, e. a. (2006). "miTarget: microRNA target gene prediction using a support vector machine." BMC Bioinformatics **7**.

Saeys, Y., et al. (2007). "A review of feature selection techniques in bioinformatics." Bioinformatics **23**(19): 2507-2517.

Salzberg, e. a. (1998). "Microbial gene identification using interpolated markov models." Nucleic Acids Research **26**: 544-548.

Scutari, M. (2010). "Learning Bayesian Networks with the bnlearn R Package." Statistical Software **53**(3): 1-22.

Shan, T. J., et al. (2009). "Application of genetic algorithm in data mining." 1st Int Work Educ Technol Comput Sci. IEEE **2**: 353-356.

Sinha, S. (2003). "Discriminative motifs." J. Comput. Biol **10**: 599-615.

Team, R. C. (2015). "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria.

Tohka, J., et al. (2016). "Comparison of Feature Selection Techniques in Machine Learning for Anatomical Brain MRI in Dementia." Neuroinform **14**: 279–296.

Tsamardinos, I., et al. (2012). "Discovering Multiple, Equivalent Biomarker Signatures." In 7th Conference of the Hellenic Society for Computational Biology and Bioinformatics.

Witten, I. H., et al. (2011). "Data mining practical machine learning tools and techniques." Morgan Kaufmann publisher, Burlington.

Xuan, P., et al. (2011). "Genetic algorithm-based efficient feature selection for classification of pre-miRNAs." Genetics and Molecular Research **10**(2): 588-603.