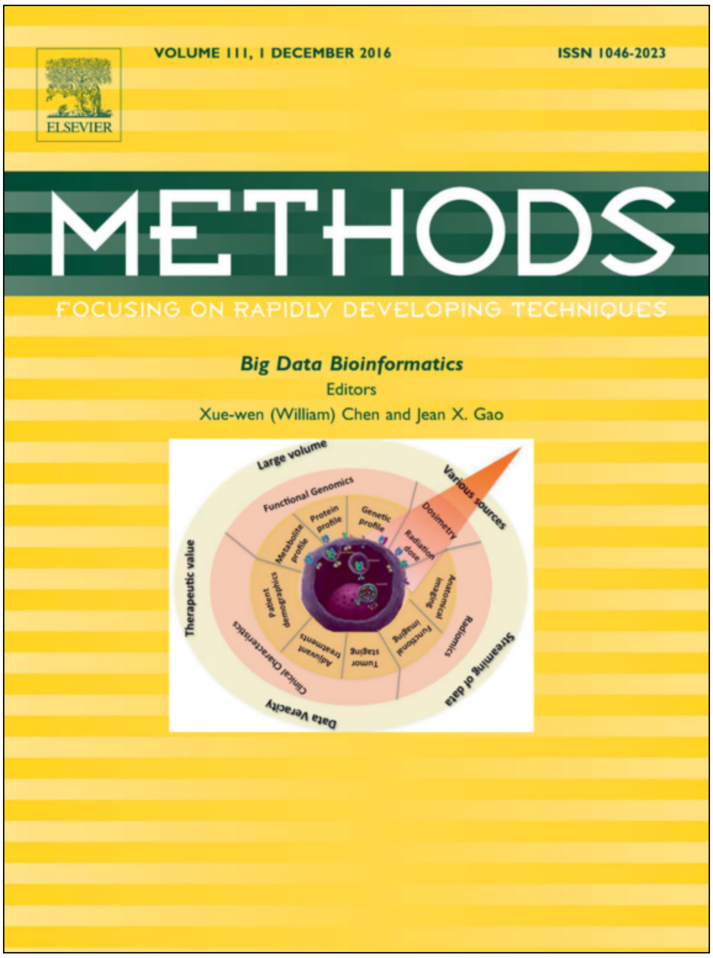

VOLUME 111, 1 DECEMBER 2016 ISSN 1046-2023

ELSEVIER

METHODS
FOCUSING ON RAPIDLY DEVELOPING TECHNIQUES

Big Data Bioinformatics
Editors
Xue-wen (William) Chen and Jean X. Gao

Methods 111 (2016) 21–31

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Feature selection methods for big data bioinformatics: A survey from the search perspective

Lipo Wang[a], Yaoli Wang[b,*], Qing Chang[b,*]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[b] College of Information Engineering, Taiyuan University of Technology, Taiyuan, China

ARTICLE INFO

Article history:
Received 15 May 2016
Received in revised form 25 August 2016
Accepted 30 August 2016
Available online 31 August 2016

Keywords:
Biomarkers
Classification
Clustering
Computational biology
Computational intelligence
Data mining
Evolutionary computation
Evolutionary algorithms
Fuzzy logic
Genetic algorithms
Machine learning
Microarray
Neural networks
Particle swarm optimization
Pattern recognition
Random forests
Rough sets
Soft computing
Swarm intelligence
Support vector machines

ABSTRACT

This paper surveys main principles of feature selection and their recent applications in big data bioinformatics. Instead of the commonly used categorization into filter, wrapper, and embedded approaches to feature selection, we formulate feature selection as a combinatorial optimization or search problem and categorize feature selection methods into exhaustive search, heuristic search, and hybrid methods, where heuristic search methods may further be categorized into those with or without data-distilled feature ranking measures.

Contents

* Corresponding authors.
E-mail addresses: elpwang@ntu.edu.sg (L. Wang), wangyaoli@tyut.edu.cn (Y. Wang), changqing@tyut.edu.cn (Q. Chang).

http://dx.doi.org/10.1016/j.ymeth.2016.08.014
1046-2023/© 2016 Elsevier Inc. All rights reserved.

1. Introduction

We have entered an era of big data. Data are becoming bigger not only in terms of the abundance of patterns (data instances or tuples), but also the dimensionality of features (or data attributes). This can significantly degrade the accuracy and efficiency of most learning algorithms, especially when there exist irrelevant or redundant features. Sometimes, the sheer size of the data even renders the data mining algorithms completely useless. The situation is particularly acute in bioinformatics [1,2,3,4,5,6,7].

Laney [8] characterized big data with 3Vs, i.e., volume (enormous size of data sets), variety (many sources and types of data) and velocity (fast pace at which data flows in from sources) [9]. Later, Normandeau [10] added the 4th V, i.e. veracity, to describe biases, noise and abnormality in data. A large variety of bioinformatics data includes genomics, proteomics, biomedical imaging, clinical trial data, etc. In particular, Stephens [11] compared genomics with three other major generators of big data: Twitter, YouTube, and astronomy, in terms of annual storage space required. Twitter requires 1–17 PB (Petabyte or 1 million GB), whereas YouTube and astronomy require 1 EB (Exabyte or 1 billion GB) and 1–2 EB, respectively. Genomics, which is only a part of bioinformatics data, requires 2–40 EB per year! The authors further estimated that between 100 million and as many as 2 billion human genomes could be sequenced by the year 2025, representing 4–5 orders of magnitude growth in 10 years and far exceeding the growth in the 3 other big data domains. Privacy concerns often require biomedical data to be modified before analysis, which can exacerbate veracity in the bioinformatics domain, compared to other big data domains.

Langley et al. [12,13] pointed out that the predictive accuracy of the learning algorithms are reduced in the presence of irrelevant features. Koller et al. [14] proved that the distribution of truly relevant features for the main task are blurred by irrelevant or redundant features [15]. Fu and Wang [16] showed that deleting those irrelevant features can not only improve the classification accuracy, but also reduce the structural complexity of the radial basis function (RBF) neural network and facilitate rule extraction. Hence data dimensionality reduction (DDR) is of paramount importance for mining big data [15,17,18,19].

There are various taxonomies for DDR. Depending on whether or not the original features are transformed into new features, one may categorize DDR techniques into feature extraction or feature selection (FS) techniques, respectively. Depending on whether or not a classifier is used to evaluate the performance of a feature subset during feature search, DDR techniques can be categorized into wrapper or filter methods, respectively. Feature extraction methods, e.g., principal component analysis (PCA) [20], linear discriminant analysis (LDA) [21,22,23] and extensions [24,25,26,27] transform the original set of features into a new set of features. Because the new features are different from the original features, it may be difficult to interpret the meaning of the new features.

There are some excellent survey papers related to soft computing techniques, FS, machine learning approaches, and bioinformatics, for example, [1,2,3,4,5,6,7]. Saeys et al. presented a review on FS techniques in bioinformatics, although this paper was published nearly a decade ago and was not so much concerned with big data. Some of the existing survey papers are for specific domain problem-solving in bioinformatics, not necessarily focusing on FS, for example, Lazar et al. [28] presented a survey on filter techniques for FS in gene expression microarray analysis. Mitra et al. [2] reviewed soft computing methodologies used in genetic networks. Phan et al. [3] discussed a biomarker identification pipeline in cardiovascular genomics. Chen et al. [4] described several intelligent techniques for identifying single nucleotide polymorphism (SNP) interactions. Kourou et al. [5] surveyed various machine learning applications in cancer prognosis and prediction. Neto [6] discussed methods for microarray classification. Other overview papers are more general, for example, Liang et al. [7] described computational functional genomics. In our present paper, we attempt to survey more recent studies on FS in bioinformatics and relate to big data whenever appropriate. We emphasize on the algorithmic aspects in FS, rather than the domain problems.

The rest of the paper is organized as follows. In Section 2, we discuss FS in general and view FS as a combinatorial optimization or search problem. We then describe truly optimal FS, i.e., exhaustive search. Section 3 reviews various sub-optimal FS approaches, i.e., heuristic search methods with and without data-distilled feature importance ranking. Section 4 covers hybrid FS approaches, i.e., semi-exhaustive search and other hybrid FS techniques. Finally, Section 5 concludes the paper and discuss future challenges.

2. Truly optimal feature selection: exhaustive search

FS aims at selecting a subset (or subsets) of the original features while achieving the best for a pre-determined objective, often the highest classification accuracy (for test data) [17]. Note that the best feature set may very well be the original features in the absence of redundant or irrelevant features. In the presence of redundant or irrelevant features, there can be multiple subsets of features that are equally good for a given objective, that is, FS may not necessarily result in a unique subset of features.

FS eliminates redundant and irrelevant features, and obtains the best subset(s) of features which most successfully discriminate(s) among classes. FS techniques are widely explored because it is easier to interpret *selected* features compared to *extracted* new features. Numerous applications, including document classification, object recognition, disease diagnosis, and computer vision, require the aid from FS.

Among the large number of classifiers available, it will be sensible to choose good classifiers, such as random forests [29], support vector machines (SVMs) [30,31,32], cluster-oriented ensemble classifiers [33], random vector functional link (RVFL) [34], and radial basis function (RBF) neural networks [35,36,37].

Searching for the truly optimal subset(s) of features is usually computationally expensive and has been shown to be NP-hard [38]. Basically, in order to find the very best feature subset(s), one would need to exhaustively try out all possible M-feature combinations of the N original features, with M=1,2, … ,N. This “combinatorial explosion” for an exhaustive search leads to a computational load that grows exponentially as the total number of features increases. In practical terms, this becomes impossible even for the most powerful computers if there are more than 30 features to be searched.

3. Sub-optimal feature selection: heuristic search

Hence, in most practical situations, exhaustively searching for the truly optimal subset(s) of features cannot be accomplished. For a combinatorial optimization problem like FS, one has no choice but resort to various types of heuristic search techniques. "Heuristic search" is search guided by "experience" or "sensible choices", "in the hope" that good sub-optima or even global optima are searched before other unfruitful subsets. In practice, well-designed heuristic search is likely to out-perform random search, although this cannot be guaranteed, since in some cases, random search could quickly stumble on a global optimum by luck. The actual performance of a heuristic search depends largely on the design of the heuristic algorithm, as well as the problem at hand, i.e., the dataset in consideration for FS.

Furthermore, good heuristic search algorithms should find a global optimum if given enough time to search.

There are many well-known heuristic search algorithms which have two essential ingredients: (i) local improvement and (2) innovation. The first ingredient "local improvement" is to search the neighbors in the state space for solutions that are better than the current solution. If an algorithm does only local improvements, for example, gradient decent or steepest ascent (depending on whether the solution evaluator is a cost function or an objective function), the algorithm may become stuck at a local optimum. The second ingredient, "innovation", allows the search to accept some solutions that are not as good as the present solution, so that the search can escape from local optimum.

For example, simulated annealing [39,40] mimics the physical annealing process where metals are heated up to a high temperature and gradually cooled down, so that the atoms move around at the high temperature and settle down at better locations as the metal cools, thereby hardening the metal. During search in simulated annealing, better solutions are always accepted, while worse solutions are accepted with a probability that reduces as the "temperature" (now a variable) decreases. The algorithm is able to find good local optima within a reasonable period of time and would always find a global optimum if annealing is carried out sufficiently slowly.

A genetic algorithm (GA) [41,42,43,44] is a heuristic search algorithm inspired from the natural evolutionary process and has been widely used in solving many optimization problems. The simplest form of GA starts with a population of randomly generated candidate solutions (phenotypes or individuals). Each individual has a fitness value according to the objective function that quantifies how good the candidate solution is. Each candidate solution has a set of properties (chromosomes or genotype) usually represented as binary strings of 0s and 1s. In each generation, the fitter individuals are selected from the current population, and each individual's genome is modified, i.e., by cross-over and randomly mutation, to form a new generation. The algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached.

Other heuristic search algorithms include ant colony optimization (ACC) [45,46], particle swarm optimization (PSO) [47,48,49], chaotic simulated annealing [50,51], tabu search [52,53], noisy chaotic simulated annealing [54,55,56,58], branch-and-bound [57], etc.

3.1. Feature selection via heuristic search without data-distilled feature importance ranking

Siedlecki et al. [59] and Raymer et al. [23] proposed an approach to selecting feature subsets using GA, where GA was used to find a binary vector (feature mask) and each bit represents the presence (1) or absence (0) of a feature. The nearest neighbor classifier was used to evaluate the fitness (the classification accuracy) of each feature subset. Xiong [60] developed a novel hybrid method of case-based reasoning and GA for FS. Case-based reasoning is carried out on a leave-one-out procedure to obtain an error estimate, which is used together with selected attributes to provide an evaluation function for GA search.

Fu and Wang used GA to select features, thereby simplifying the structure of a RBF neural network and extracting succinct rules from data [61,62,63]. Lin et al. [64] used GA for FS and proposed silhouette statistics as a discriminant function to distinguish between six subtypes of pediatric acute lymphoblastic leukemia by using microarray with thousands of gene expressions. Kleftogiannis et al. [65] combined SVM with genetic algorithms (GA) for FS and parameters optimization. The best model trained on human data successfully predicted pre-miRNAs to other organisms including the category of viruses.

Zhang [66] introduced peak tree to represent the peak information in mass spectrometry (MS) spectra. They presented an improved ant colony optimization biomarker selection method to build the MS analysis system. Experiments on real SELDI data showed that their peak detection method can better resist spectrum variations and provide higher sensitivity and lower false detection rates than other methods.

After a good classifier, such as a SVM or a neural network, is trained, the input weights for any irrelevant input features should become very small or vanish. One can then eliminate those input features associated with small weights in the classifier trained with the data. This is sometimes called the "embedded" FS method, i.e., classifiers embedded with FS procedures. If one desires, one can still call these weights in the classifiers some sort of feature importance ranking measures; however, these feature importance measures are derived from trained classifiers, rather than distilled directly from input data (as in the approaches to be covered in the next subsection). Similarly, for a regression problem, e.g., risk prognosis and time series prediction, a regression algorithm (e.g., support vector regression or SVR) should be used as the objective evaluator, in place of a classifier. Occasionally, instead of classifiers, clustering algorithms can also be used. These "embedded" FS methods are sub-optimal, since training of the objective evaluator (classifier, regression algorithm) is not guaranteed to reach global optima. All methods below in this subsection belong to "embedded" FS.

Guyon et al. [67] proposed a support vector machine method based on recursive feature elimination (RFE). The SVM-RFE algorithm trains a SVM with a linear kernel and removes the feature with smallest w value of the decision hyperplane given by the trained SVM. Zhong et al. [68] used SVM-RFE to select features for predicting essential proteins and removed the features that share biological meaning with other features according to Pearson Correlation Coefficients (PCC). Experiments were carried out on S. cerevisiae data. 6 features were determined as the best subset. They compared with SVM, Naive Bayes, Bayes Network, and NBTree.

Furlanello et al. [69] presented E-RFE, a variant of the SVM-RFE in which features are eliminated using entropy of the weights distribution of a SVM classifier. A dynamic time warping (DTW) algorithm was then applied to define a metric between sample-tracking profiles. An unsupervised clustering based on the DTW metric allowed for automating the discovery of outliers and of subtypes of different molecular profiles. They applied the algorithm to synthetic data and two gene expression studies.

Duan et al. [70] carried out a statistical analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data, as opposed to one such SVM in SVM-RFE. They tested the method on four gene expression datasets for cancer classification. This study suggested that, for gene expression-based cancer classification, an average test error from multiple partitions of training and test sets can be recommended as a reference of performance quality.

To address the problem that the SVM-RFE is highly sensitive to the "filter-out" factor, i.e., how many genes are removed at one step, Tang et al. [71] proposed a two-stage SVM-RFE algorithm for microarray gene selection. The algorithm effectively eliminated most of the irrelevant, redundant and noisy genes while keeping information loss small at the first stage. A fine selection for the final gene subset was then performed at the second stage. The two-stage SVM-RFE overcame the instability problem of the SVM-RFE and achieved improved results.

Yousef et al. [72] proposed the SVM-RCE technique, which combines K-means, a clustering method, to identify correlated gene clusters, and the SVM, a supervised machine learning classification method, to identify and score (rank) those gene clusters for the purpose of classification. K-means is used initially to group genes into clusters. Recursive cluster elimination (RCE) is applied to iteratively remove those clusters of genes that contribute the least to classification. Luo et al. [73] improved recursive cluster elimination based on SVM (SVM-RCE). This ISVM-RCE method first trains a SVM model with all clusters, then applies the infinite norm of weight coefficient vector in each cluster to score the cluster, finally eliminates the gene clusters with the lowest score. In addition, ISVM-RCE eliminates genes within the clusters instead of removing a cluster of genes when the number of clusters is small. The authors tested ISVM-RCE on six gene expression data sets and compared their performances with the original SVM-RCE and lin ear-discriminant-analysis-based RFE (LDA-RFE). The experiment results on these data sets show that ISVM-RCE greatly reduces the time cost of SVM-RCE, while obtaining comparable classification performance as SVM-RCE.

Li et al. [74] designed margin influence analysis (MIA) to work with SVM for selecting informative genes. MIA should reveal genes which have statistically significant influence on the margin in the SVM by using Mann–Whitney *U* test. The main reason for using the Mann–Whitney *U* test rather than two-sample t test is that the former is a nonparametric test method without any distribution-related assumptions.

To predict protein structure classes, Hayat et al. [75] proposed a model employing a hybrid descriptor space in conjunction with an optimized evidence-theoretic K-nearest neighbor algorithm. The hybrid space consists of two descriptor spaces including multi-profile Bayes and bi-gram probability. The authors selected discriminative descriptors from the hybrid space using PSO. They compared the performance of their approach to other algorithms on benchmarking data.

Blocking is an experimental design strategy where similar conditions are used to compare alternative configurations to assure that observed differences in accuracy are due to underlying differences rather than to fluctuations or noise. Bontempi [76] proposed a blocking strategy for improving FS which aggregates the validation outcomes of several learning algorithms to assess a gene subset. The paper showed that the blocking strategy significantly improves the performance of a conventional forward selection for a set of 16 publicly available cancer expression data sets. The experiments involve six different classifiers and show that improvements are independent of the classification algorithm used.

Quantitative structure–activity relationships (QSARs) correlate biological activities of chemical compounds with their physicochemical descriptors. Wong and Burkowski [77] used kernel alignment as an evaluation tool, using recursive feature elimination to compute a molecular descriptor containing the most important features needed for a classification application. Their empirical results showed that the algorithm worked well for the computation of descriptors for various applications involving different QSAR data sets.

To automatically detect lexico-semantic event structures in biomedical texts, Ozyurt [78] designed a noun argument structure (NAS) annotated corpus and a SRL system to identify and classify these structures. A GA-based FS (GAFS) method was used to significantly improve the performance of the system.

Ghorai et al. [79] carried out GA-based simultaneous feature and model selection for an ensemble of nonparallel plane proximal classifiers (NPPCs). Experimental results on cancer data sets showed that the NPPC ensemble offers comparable testing accuracy to that of SVM ensembles with reduced training time on average.

Fong et al. [80] used PSO to search for optimal feature subsets, together with three classifiers, i.e., pattern network, decision tree, and Navies Bayes. The approach was shown to achieve high accuracy in classification in two empirical biomedical datasets, i.e., the Arrhythmia dataset and the MicroMass dataset.

Sun et al. [81] extended the L1-L2 SVM classifier proposed by others [82,83,84] to regression analysis, thereby achieving cancer prognosis and automatic FS from the trained SVR. The proposed method was compared with other seven prognostic prediction methods on three real-world data sets.

Liu et al. [85] proposed a SVMs with L_p ($p < 1$) regularization that is applicable to deal with high-dimensional data sets with both discrete and continuous data types. The regularization parameters were estimated through maximizing the area under the ROC curve (AUC) of the cross-validation data. They carried out experiments on protein sequence and SNP data. The accurate and sparse L_pSVM leads to effective FS.

The support feature machine (SFM) finds the least number of features of a data set such that two classes are linearly separable without error. The dimensionality of the data is thus more efficiently reduced than with support vector based FS. Klement et al. [86] provided a new formulation of the SFM where classification of unbalanced and nonseparable data was straightforward. They applied the SFM to a functional magnetic resonance imaging data set.

Mohapatra et al. [87] proposed two variations of kernel ridge regression (KRR) [88,89,90], namely wavelet kernel ridge regression (WKRR) and radial basis kernel ridge regression (RKRR) for classification of microarray medical datasets. The authors also proposed a modified cat swarm optimization (MCSO), a naturally inspired evolutionary algorithm, to select the most relevant features from the datasets. They used several biomedical datasets to demonstrate their algorithms.

Maji and Pal [91] proposed a rough-fuzzy C-Medoids algorithm for selection of bio-basis for amino acid sequence analysis. The membership function of fuzzy sets enables efficient handling of overlapping partitions, while lower and upper bounds of rough sets [92,93] deal with uncertainty, vagueness, and incompleteness in class definition. The crisp lower bound and fuzzy boundary of a class in the algorithm enables efficient selection of the minimum set of the most informative bio-bases. Maji and Partha Garai [94] presented a FS method based on fuzzy-rough sets by maximizing both relevance and significance of the selected features. They also presented different feature evaluation criteria such as dependency, relevance, redundancy, and significance for attribute selection using fuzzy-rough sets. The performance of different rough set models was compared with some existing techniques based on the accuracy of nearest neighbor rule, SVM, and decision tree on a set of microarray gene expression datasets.

Maulik and Chakraborty [95] proposed a fuzzy preference based rough set (FPRS) method for feature (gene) selection with semisupervised SVMs. They compared the performance of this technique with the signal-to-noise ratio (SNR) and consistency based FS (CBFS) methods. They demonstrated their scheme using six benchmark gene microarray datasets, including both binary and multi-class classification problems.

Pang et al. [96] used the random forests to identify a set of prognostic genes. They compared their method with several machine learning methods and various node split criteria using several real data sets. They showed that their method incorporates multivariate correlations and is advantageous over single-gene-based approaches.

Wu et al. [97] proposed a Laplace naive Bayes model with mean shrinkage (LNB-MS). The Laplace distribution was used instead of the normal distribution as the conditional distribution of the samples because it is less sensitive to outliers. The L_1 penalty was imposed on the mean of each class to achieve automatic FS. Experimental results were shown for simulated data sets and 17 publicly available cancer data sets.

Array comparative genomic hybridization (aCGH) is a relatively new method for the detection of copy number abnormalities associated with human diseases with special focus on cancer. Metsis et al. [98] presented a FS method based on structured sparsity-inducing norms to identify the informative aCGH biomarkers. They experimentally compared the proposed approach with existing FS methods on four publicly available aCGH data sets.

Boareto et al. [99] introduced an analytic geometric FS method called supervised variational relevance learning (Suvrel), a variational method to determine metric tensors to define distance based similarity in pattern classification. The variational method was applied to a cost function that penalizes large intraclass distances and favors small interclass distances. They found analytically the metric tensor that minimizes the cost function.

Tan et al. [100] proposed a minimax sparse logistic regression (LR) model for very high-dimensional FSs, which can be efficiently solved by a cutting plane algorithm. To solve the resultant nonsmooth minimax subproblems, the authors designed a smoothing coordinate descent method.

Wang et al. [101] designed unified objectives for FS, multiple kernel learning, sparse coding, and graph regularization. By optimizing the objective functions iteratively, they achieved FS and multiple kernel learning. They demonstrated their approach with experimental results on two challenging tasks, N-linked glycosylation prediction and mammogram retrieval.

Garcia-Pedrajas et al. [102] proposed an evolutionary simultaneous instance and FS algorithm that is scalable to millions of instances and thousands of features, using divide-and-conquer. They demonstrated their algorithm with 13 very large datasets.

3.2. Greedy search with data-distilled feature importance ranking

Although FS approaches via heuristic search discussed in the previous subsection are less computationally expensive compared to exhaustive search, it may still take a long time to find a good feature subset. An alternative approach is to first evaluate the "importance" of each feature individually, using some sort of feature importance information distilled from the data, and then either select a certain number of the most important features without any search aided by a classifier (often called the filter approach [103]), or by simple greedy search with the help of a classifier (the wrapper approach). A greedy search can be either a backward search (the least important features are gradually removed from the full feature set) [104] or forward search (the most important features are gradually added to an empty feature set) [105].

Even for the filter methods, very often the final goal is classification and a classifier is eventually involved. For example, if an input feature represents a biomarker for a disease, this input feature should contribute positively towards classification (diagnosis) performance. Although there are numerous types of classifiers and a feature subset best for one classifier may not be the best for another classifier, it is not a good idea to leave out classifiers altogether for FS.

Examples of data-distilled feature importance measures include *t*-test [106,107], fold-change difference [108], Z-score [109], Pearson correlation coefficient [110], relative entropy [111], mutual information [112,113], separability-correlation measure [16], feature relevance [114,115], label changes produced by each feature [116], information gain [117], etc. The feature importance is directly derived from the input data, as opposed to being extracted from a trained classifier.

Chu and Wang [118,119] used the SVM for cancer classification with microarray data. Dimensionality reduction methods, such as class-separability measure, Fisher ratio, principal components analysis (PCA), and *t*-test, were used for gene selection. A voting scheme was then employed to do multi-group classification by multiple binary SVMs. They were able to obtain the same classification accuracy but with much fewer features compared to other published results.

Liu et al. [120] carried out comparisons on FS for Affymetrix (Affy) microarray studies across different labs. They investigated four FS methods: *t*-test, significance analysis of microarrays (SAM), rank products (RP), and random forest (RF). They applied the four methods to acute lymphoblastic leukemia, acute myeloid leukemia, breast cancer, and lung cancer Affy data which consist of three cross-lab data sets each. Their results showed that SAM has the best classification performance. RF also obtained high classification accuracy, but was not as stable as SAM.

Zhou and Wang [121,122] first ranked each feature (SNP) using a modified *t*-test or *F*-statistics. From the ranking list, they formed different feature subsets by sequentially choosing different numbers of features (e.g., 1, 2, 3, . . ., 100.) with top ranking values, train and test them by the SVM, thereby finding one subset which had the highest classification accuracy. Their method was able to identify a very small number of important SNPs that can determine the populations of individuals.

To address the problem that the selected genes by the same method often vary significantly with sample variations, Yu et al. [123] proposed a general framework of sample weighting to improve the stability of FS methods under sample variations. The framework first weights each sample in a given training set according to its influence to the estimation of feature relevance, and then provides the weighted training set to a FS method. Experiments on a set of microarray data sets showed that the proposed algorithm significantly improved the stability of representative FS algorithms such as SVM-RFE and ReliefF, without sacrificing their classification performance.

Peters et al. [124] used cross-entropy on a human lymph node data set to show that a significant number of genes perform well when their complementary power is assessed, but "pass under the radar" of popular FS methods that only assess genes individually on a given classification tool. They also showed that this phenomenon becomes more apparent as diagnostic specificity of the tissue samples analyzed increases.

Valavanis, Maglogiannis, and Chatziioannou [125] used an integrated multimodal dataset related to cutaneous melanoma that fuses two separate sets providing complementary information, i.e., gene expression profiling and imaging. The first goal was to select a subset of genes that comprise candidate genetic biomarkers. The derived gene signature was then utilized in order to select imaging features, which characterize disease at a macroscopic level, presenting mutual information content to the selected genes. Using information gain ratio and exploration of the gene ontology tree, they identified a set of 32 uncorrelated genes significant to melanoma. Selected genes and imaging features were used to train various classifiers that could generalize well when discriminating malignant from benign melanoma samples.

Gumus et al. [126] used a multi objective optimization technique called Pareto Optimal for selecting SNP subsets offering both high classification accuracy and correlation between genomic and geographical distances. Discriminatory power of an SNP was determined using mutual information. They demonstrated their method with the Human Genome Diversity Project (HGDP) SNP dataset.

Bennasar et al. [127] introduced Joint Mutual Information Maximization (JMIM) and Normalized Joint Mutual Information Maximization (NJMIM); both these methods use mutual information and the maximum-of-the-minimum criterion, which alleviates the problem of overestimation of the feature significance as demonstrated both theoretically and experimentally. They compared their proposed methods using eleven publicly available datasets with five competing methods.

Maji [128] presented different f-information measures as the evaluation criteria for gene selection. To compute the gene-gene redundancy, these information measures calculate the divergence of the joint distribution of two-gene expression values from the joint distribution when two genes are considered to be completely independent. The performance of different f-information measures was compared with that of mutual information based on accuracy of naive Bayes classifier, K-nearest neighbor rule, and SVM on the breast cancer, leukemia, and colon cancer datasets.

Ranganarayanan et al. [129] used information gain and SVM to select glucose-binding pockets in human serum albumin (HSA). The predictions were further corroborated using docking studies. They argued that these findings can complement studies directed towards the development of HSA as an alternate biomarker for glycemic monitoring.

Leung et al. [130] selected potential markers for DNA sequences of Hepatitis B virus based on information gain for further classifier learning. They developed a new classification method by nonlinear integral and collected HBV DNA sequences from over 200 patients specifically for this project.

Phosphorylation is a crucial post-translational modification and regulates most cellular processes. Xu et al. [131] used minimum-redundancy-maximum-relevance (mRMR) measure [132] and forward FS, together with the SVM. Their results outperformed other classifiers, such as Bayesian decision theory, k nearest neighbor and the random forest.

Mohabatkar et al. [133] implemented the mRMR FS method, together with the SVM, and successfully predicted amino acid gamma-aminobutyric-acid receptors (GABAARs).

To address the problem that the original mRMR FS model for minimizing the redundancy between sequentially selected features uses a greedy search, Wang et al. [134] attempted to globally minimize the feature redundancy by maximizing the given feature ranking scores, which can come from any supervised or unsupervised methods.

Lin et al. [135] added combinatorial fusion to their previous hierarchical learning architecture (HLA) using neural networks for protein structure prediction. Feature selection was facilitated with a diversity score function. The resulting classification has an overall prediction accuracy rate of 87% for four classes and 69.6% for 27 folding categories, which were significantly higher than the accuracy rate of 56.5% previously obtained by Ding and Dubchak.

To identify main trends of activity through time in gene time series, Furlanello et al. [136] proposed a reconstruction method based on stagewise boosting, using a similarity measure based on the dynamic time warping (DTW) algorithm and defining a ranked set of time-series components contributing most to the reconstruction. The approach was applied on synthetic and public microarray data.

Mohammadi et al. [137] proposed a Maximum-Minimum Correntropy Criterion (MMCC) for selection of informative genes from microarray data sets. They determined the optimal number of features for each dataset using evolutionary optimization. They showed that their algorithm worked better compared to other algorithms for 25 commonly used microarray data sets. In particular, they showed that high accuracy in classification by SVM was achieved by fewer than 10 genes in all of the cases.

To model prior knowledge about the network topology in the inference problem of gene regulatory networks (GRNs) from expression profiles, Lopes et al. [138] aggregated scale-free properties to the Sequential Floating Forward Selection (SFFS) FS method to guide the inference. They carried out experiments using synthetic and real data, and showed that this technique provided smaller estimation errors compared to techniques without aggregating scale-free properties.

Zhang et al. [139] proposed an unsupervised feature ranking method to evaluate the importance of the features based on consensus affinity, by comparing the corresponding affinity of each feature between a pair of instances based on the consensus matrix of clustering solutions. Experiments on real gene expression data sets demonstrated improvement over other techniques.

4. Hybrid feature selection techniques

4.1. Semi-exhaustive search

Wang et al. [140] attempted at finding the smallest set of genes that can ensure highly accurate classification of cancers from microarray data. Their simple yet effective approach consists of the following two steps. In the first step, some important genes are chosen using a feature importance ranking measure. In the second step, an exhaustive search is carried out within the top-ranked genes: the classification capability of each simple (i.e., 1-gene, 2-gene, 3-gene) combinations of those important genes is evaluated using a good classifier. For three "small" and "simple" data sets with 2, 3, and 4 cancer (sub) types, this approach leads to very high accuracy with only 2 or 3 genes. For a "large" and "complex" data set with 14 cancer types, the authors divided the whole problem into a group of binary classification problems and applied the 2-step approach to each of these binary classification problems. Through this "divide-and-conquer" scheme, they obtained accuracy comparable with previously reported results *but with only 28 genes rather than 16063 genes*. In general, this semi-exhaustive search method can significantly reduce the number of genes required for highly reliable diagnosis.

In another study, Li and Yin [141] proposed a multi-objective biogeography based optimization method to select small subsets of informative genes. They first used the Fisher-Markov selector choose the 60 top genes. Multi-objective binary biogeography based optimization (MOBBBO) was proposed for gene selection and support vector machine was used as the classifier with the leave-one-out cross-validation method (LOOCV).

Wu et al. [142] propose a stratified sampling method for feature subspace selection to generate decision trees in a random forest for genome-wide association (GWA) high-dimensional data. To avoid the high computational costs associated with exhaustive search, they devised an equal-width discretization scheme to divide SNPs into multiple groups. They randomly selected the same number of SNPs from each group and combined them to form a subspace to generate a decision tree. The method was demonstrated using two genome-wide SNP data sets, i.e., the Parkinson case-control data (408 803 SNPs) and the Alzheimer case-control data (380 157 SNPs).

Bonilla-Huerta et al. [143] proposed a hybrid framework composed of two stages for gene selection and classification of DNA microarray data. At the first stage, five traditional statistical methods were combined for preliminary gene selection (Multiple Fusion Filter). Then different relevant gene subsets were selected by using an embedded GA, Tabu Search (TS), and SVM. A gene subset was obtained by analyzing the frequency of each gene in the different gene subsets. The most frequent genes were shown to be a final gene subset with high performance.

Sajjadi et al. [144] proposed a network-based framework to identify effective biomarkers by searching for groups of synergistic risk factors with high predictive power to disease outcome. They constructed an interaction network with node weights representing individual predictive power of candidate factors and edge weights capturing pairwise synergistic interactions among factors. They then formulated this network-based biomarker identification problem as a graph optimization model to search for multiple cliques with maximum overall weight. To search for optimal or near optimal solutions, both an analytical algorithm based on column generation method and a fast heuristic for large-scale networks were derived. They demonstrated their algorithms with two biomedical data sets: a Type 1 Diabetes (T1D) data set from the Diabetes Prevention Trial-Type 1 (DPT-1) study and a breast cancer genomics data set for metastasis prognosis.

4.2. Other hybrid feature selection approaches

Feature extraction techniques can be used to effectively aid FS.

Liu et al. [145] proposed an efficient method for selecting relevant genes. Firstly they used spectral biclustering to obtain the best two eigenvectors for class partition. Then gene combinations were selected based on the similarity between the genes and the best eigenvectors. They demonstrated their semi-unsupervised gene selection method using two microarray cancer data sets, i.e., the lymphoma and the liver cancer data sets, where their method was able to identify a single gene or a two-gene combinations which can lead to predictions with very high accuracy.

In order to improve the performance of sparse principal component analysis, Liu et al. [146] proposed a class-information-based sparse component analysis (CISCA) method which introduces the class information via a total scatter matrix. They first normalized the RNA-Seq data by using a Poisson model to obtain their differential sections. They then obtained the total scatter matrix by combining the between-class and within-class scatter matrices. Third, they decomposed the total scatter matrix by using singular value decomposition and constructed a new data matrix by using singular values and left singular vectors. To obtain sparse components, they further decomposed the constructed data matrix by solving an optimization problem with sparse constraints on loading vectors. Finally, they identified differentially expressed genes by using the sparse loading vectors. The method was demonstrated with results on real RNA-Seq data.

Pinto da Costa et al. [147] proposed a weighted PCA-based FS method, by pointing out that the principle components are weighted sums of various features, which indicates the importance of each features. They applied the method, together with SVM, to microarray data.

Liu et al. [148] used robust principal component analysis (RPCA) and linear discriminant analysis (LDA) to identify the features of gene expression data. The SVM was applied to classify the tumor samples of gene expression data based on the identified features.

Niijima and Okuno [149] extended Laplacian linear discriminant analysis (LLDA) to unsupervised cases and proposed an unsupervised FS method, called LLDA-based recursive feature elimination (LLDA-RFE). They applied LLDA-RFE to several public data sets of cancer microarrays and compared its performance with those of Laplacian score and SVD-entropy, two state-of-the-art unsupervised methods, and with that of Fisher score, a supervised filter method.

Zheng et al. [150] select genes using nonnegative matrix factorization (NMF) or sparse NMF (SNMF) and then extracted features from the selected genes by virtue of NMF or SNMF. They used SVM to classify the tumor samples using the extracted features.

Considering the nonstationary characteristics of surface electromyography (sEMG), signals of superficial muscles from the skin surface, Naik and Nguyen [151] use NMF to select features for hand gesture recognition. They conducted experiments for simple and complex finger flexions with an artificial neural network classification scheme.

5. Conclusions and future challenges

We have surveyed main principles of FS and their recent applications in big data bioinformatics. Instead of the commonly used categorization into filter, wrapper, and embedded FS approaches, we viewed FS as a combinatorial (discrete) optimization problem and categorize FS methods into exhaustive search, heuristic search, and hybrid methods, where heuristic search methods may further be categorized into those with or without data-distilled feature ranking measures.

Tremendous amount of excellent research has been produced in this area and major progress has been made. Some challenges still remain.

5.1. The small sample size problem

In some biomedical problems, notably for DNA microarray data, the dimensionality can be quite high (e.g., up to 20,000 genes), whereas the sample size is rather small (e.g., around 50 patients) [152]. In this situation, the number of independent variables exceeds by far the number of training samples, leading to possible overfitting and overoptimism.

For example, Wang et al. [140] found that a microarray data with thousands of genes can be classified with 100% accuracy with only 2 genes and there are many 2-gene combinations that can do this. Adequate cautions were taken during training to assure that testing data were not involved in FS during training. From a pure computational point of view, all these 2-gene combinations are equally good feature subsets; however, 2 important questions may be raised: (1) what happens if more patient data are added this the dataset? (2) Are all or any of these many genes selected truly biomarkers? Some authors used ensembles of FS methods to tackle the second question: they used a set of different FS methods for the same data and selected the final feature subset by some sort of averaging or majority vote by these different FS methods [153,154]. Further research is needed on this topic.

5.2. Imbalanced data

Most of the time, data available in different classes (e.g., control subjects and patients) are different in numbers, which is called imbalanced data or skewed data. Training results may become biased towards classes with more training data and therefore may become unreliable. There have been various ways of dealing with this problem, for example, up-sampling classes with fewer data, down-sampling classes with more data, or making classification errors sensitive to classes (cost-sensitive learning) [155,156,157].

Wasikowski and Chen [158] carried out systematic comparisons of three types of methods developed for imbalanced data classification problems and seven FS metrics evaluated on small sample data sets from different applications. They evaluated the performance of these metrics using area under the receiver operating characteristic (AUC) and area under the precision-recall curve (PRC). They compared each metric on the average performance across all problems and on the likelihood of a metric yielding the best performance on a specific problem. They showed that signal-to-noise correlation coefficient (S2N) and Feature Assessment by Sliding Thresholds (FAST) are great candidates for FS in most applications.

Zhu et al. pointed out [159] that traditional gene selection based on empirical mutual information suffers from the data sparseness issue due to the small number of samples. To overcome this, they proposed a model-based approach to estimate the entropy of class variables on the model, instead of on the data themselves. They used multivariate normal distributions to fit the data, because multivariate normal distributions have maximum entropy among all real-valued distributions with a specified mean and standard deviation. They carried out experiments on seven gene data sets and compared with other five approaches.

5.3. Class-dependent feature selection

In the above approaches to FS, one usually chooses the same feature subset for all classes in a given classification problem, which is called *class-independent* FS [16,114,23116,59]. In contrast, one may also allow for a different feature subset for each class, which is called *class-dependent* FS [160,161], since different features may have different capabilities in discriminating different classes. *Class-independent* FS commonly practiced can therefore be considered as a special case of the more general *class-dependent* FS.

Oh et al. proposed [160,161] selected class-dependent features for handwriting digits. They used the estimated class distributions to calculate each feature's class separation for the 10 digits (classes). Then in terms of class separation, an ordered list of features was provided for each class and according to the ranking list each class obtained a feature vector with a predefined dimension 256. Although all the 10 feature vectors have the same dimension, they have different feature compositions.

In [162], Fu and Wang used GA to select a feature subset for each class based on an RBF classifier. This approach made explicit use of the clustering property of the RBF neural network and therefore may not work for other types of classifiers, for example, the SVM and the multi-layer perceptron (MLP) neural network.

Baggentstoss [163,164] proposed to select class-specific features on the basis of the probability density function (PDF) projection theorem. Baggenstoss [163,164] also provided theoretical proof for this method and demonstrated applications on signal processing problems.

In addition, class-dependent feature extraction methods have been proposed. For example, Liu et al. [165] proposed to extract class-specific features through principle component analysis (PCA) from class-specific subspaces.

Wang et al. [166] proposed a general approach to class-dependent FS by arguing that for a C-class classification problem, C 2-class classifiers, each of which discriminating one class from the other classes and having a characteristic input feature subset, should in general out-perform, or at least match the performance of, a C-class classifier with one single input feature subset. For each class, they selected the best feature subset which leads to the lowest classification error rate for this class using a classifier for a given feature subset searchranking algorithm. The performance of the method was evaluated on two artificial data sets and several real-world benchmark data sets, with the SVM as the classifier, and with the RELIEF, class separability, and minimal-redundancy-maxi mal-relevancy as feature importance measures. The experimental results indicated that the *class-dependent* feature subsets found by this approach can effectively remove irrelevant or redundant features, while maintaining or improving (sometimes substantially) the classification accuracy, in comparison with other (especially class-independent) FS methods.

Zhu et al. [167] introduced full class relevant (FCR) and partial class relevant (PCR) features. Particularly, FCR denotes genes that serve as candidate biomarkers for discriminating all cancer types, whereas PCR are genes that distinguish subsets of cancer types. A Markov blanket embedded memetic algorithm was proposed for the simultaneous identification of both FCR and PCR genes. Results obtained on commonly used synthetic and real-world microarray data sets showed that the identification of both FCR and PCR genes improved classification accuracy on many microarray data sets.

Rajapakse and Mundra [168] decomposed multiclass ranking statistics into class-specific statistics and then used Pareto-front analysis for selection of genes. This alleviates the bias induced by class intrinsic characteristics of dominating classes. They demonstrated the use of Pareto-front analysis on F-score and KW-score. A significant improvement in classification performance and reduction in redundancy among top-ranked genes were achieved in experiments with both synthetic and real-benchmark data sets.

Freeman et al. [169] presented a method for multiclass classification that simultaneously formulates a binary tree of simpler classification subproblems and performs FS for the individual classifiers. The feature selected hierarchical classifier (FSHC) was tested against several well-known techniques for multiclass division. Tests were run on nine different real data sets and one artificial data set using a SVM classifier. The results showed that the accuracy obtained by the FSHC is comparable with other common multiclass SVM methods. Furthermore, the results demonstrated that the algorithm creates solutions with fewer classifiers, fewer features, and a shorter testing time than the other SVM multiclass extensions.

Since class-dependent FS requires determination of feature subsets of every class, it is likely to be more computationally expensive compared to other class-independent FS methods. However, the extra computational cost may be worthwhile in applications where improvements of accuracy or reduction of data dimensionality are crucial.

Acknowledgement

We would like to sincerely thank the reviewers for their constructive comments and suggestions that helped to significantly improve the manuscript.

References

[1] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.

[2] S. Mitra, R. Das, Y. Hayashi, Genetic networks and soft computing, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (1) (2011) 94–107, http://dx.doi.org/10.1109/TCBB.2009.39.

[3] J.H. Phan, C.F. Quo, M.D. Wang, Cardiovascular genomics: a biomarker identification pipeline, IEEE Trans. Inf. Technol. Biomed. 16 (5) (2012) 809–822, http://dx.doi.org/10.1109/TITB.2012.2199570.

[4] C.C.M. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, P. Macrossan, Methods for identifying SNP interactions: a review on variations of logic regression, random forest and bayesian logistic regression, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (6) (2011) 1580–1591, http://dx.doi.org/10.1109/TCBB.2011.46.

[5] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17, http://dx.doi.org/10.1016/j.csbj.2014.11.005. URL http://www.sciencedirect.com/science/article/pii/S2001037014000464.

[6] U.B. Neto, Fads and fallacies in the name of small-sample microarray classification – a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing, IEEE Signal Process. Mag. 24 (1) (2007) 91–99, http://dx.doi.org/10.1109/MSP.2007.273062.

[7] M.P. Liang, O.G. Troyanskaya, A. Laederach, D.L. Brutlag, R.B. Altman, Computational functional genomics, IEEE Signal Process. Mag. 21 (6) (2004) 62–69, http://dx.doi.org/10.1109/MSP.2004.1359143.

[8] D. Laney, 3-d data management: Controlling data volume, velocity and variety, Application Delivery Strategies, META Group 6 February. doi: http://goo.gl/wH3qG.

[9] M. May, Big biological impacts from big data, Science. doi:10.1126/science.opms.p1400086.

[10] K. Normandeau, Beyond volume, variety and velocity is the issue of big data veracity, Inside Big Data. doi: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.

[11] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, R. Iyer, M.C. Schatz, S. Sinha, G.E. Robinson, Big data: astronomical or genomical?, PLoS Biol 13 (7) (2015) e1002195, http://dx.doi.org/10.1371/journal.pbio.1002195.

[12] P. Langley, S. Sage, Induction of selective bayesian classifiers, in: the Tenth Conference on Uncertainty in Artificial Intelligence, 1994, pp. 399–406.

[13] P. Langley, Elements of Machine Learning, Morgan Kaufmann, 1996.

[14] D. Koller, M. Sahami, Toward optimal feature selection, in: the 13th International Conference on Machine Learning (ML), 1996, pp. 284–292.

[15] H. George John, Enhancing Learning using Feature and Example Selection (Master thesis), Department of Computer Science, Texas A&M University, 2003.

[16] X.J. Fu, L.P. Wang, Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance, IEEE Trans. Syst. Man Cybern. Part B Cybern 33 (2003) 399–400.

[17] L.P. Wang, X.J. Fu, Data Mining with Computational Intelligence, Springer-Verlag, 2005.

[18] S. Halgamuge, L.P. Wang (Eds.), Classification and Clustering for Knowledge Discovery, Springer, 2005.

[19] S. Halgamuge, L.P. Wang (Eds.), Computational Intelligence for Modeling and Predictions, Springer, 2005.

[20] L.H. Chen, S. Chang, An adoptive learning algorithm for principle component analysis, IEEE Trans. Neural Networks 6 (1995) 1255–1263.

[21] W. Malina, Two-parameter fisher criteria, IEEE Trans. Syst. Man Cybern. Part B Cybern. 31 (2001) 629–636.

[22] S. Mika, G. Ratsch, J. Weston, S.B., K.R. Mullers, Fisher discriminant analysis with kernels, in: Neural Networks Signal Processing IX 1999, 1999, pp. 41–48.

[23] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, A.K. Jain, Dimensionality reduction using genetic algorithms, IEEE Trans. Evol. Comput. 4 (2000) 164–171.

[24] L. Zhang, L.P. Wang, W. Lin, Conjunctive patches subspace learning with side information for collaborative image retrieval, IEEE Trans. Image Process. 21 (2012) 3707–3720.

[25] L. Zhang, L.P. Wang, W. Lin, Semi-supervised biased maximum margin analysis for interactive image retrieval, IEEE Trans. Image Process. 21 (2012) 2294–2308.

[26] L. Zhang, L.P. Wang, W. Lin, Generalized biased discriminant analysis for content-based image retrieval, IEEE Trans. Syst Man Cybern. Part B: Cybern. 42 (2012) 282–290.

[27] L. Zhang, L.P. Wang, W. Lin, S. Yan, Geometric optimum experimental design for collaborative image retrieval, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 346–359.

[28] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, A. Nowe, A survey on filter techniques for feature selection in gene expression microarray analysis, Bioinformatics 9 (2012) 1106–1119.

[29] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[30] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273, http://dx.doi.org/10.1007/BF00994018.

[31] L.P. Wang, B. Liu, C. Wan, Classification using support vector machines with graded resolution, IEEE International Conference on Granular Computing 2, 2005, pp. 666–670.

[32] N.K. Alham, M. Li, Y. Liu, M. Qi, A distributed SVM ensemble for large scale image classification and annotation, Comput. Math. Appl. 66 (2013) 1920–1934, http://dx.doi.org/10.1007/BF00994018.

[33] B. Verma, A. Rahman, Cluster-oriented ensemble classifier: impact of multicluster characterization on ensemble classifier learning, IEEE Trans. Knowl. Data Eng. 24 (4) (2012) 605–618, http://dx.doi.org/10.1109/TKDE.2011.28.

[34] Y. Pao, G.H. Park, D.J. Sobajic, Learning and generalization characteristics of random vector functional-link net, Neurocomputing 6 (1994) 163–180.

[35] D.S. Broomhead, D. Lowe, Multivariable functional interpolation and adaptive networks, Complex Syst. 2 (1988) 321–355.

[36] X. Fu, L.P. Wang, Linguistic rule extraction from a simplified RBF neural network, Comput. Stat. 16 (3) (2001) 361–372.

[37] L.P. Wang, X. Fu, A simple rule extraction method using a compact RBF neural network, 2nd International Symposium on Neural Networks (ISNN 2005), LNCS 3496 (2005) 682–687.

[38] J. Bins, B.A. Draper, Feature selection from huge feature sets, Eighth IEEE Int. Conf. Comput. Vision 2 (2001) 159–165.

[39] S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.

[40] L.P. Wang, N.S.L. Sally, W.Y. Hing, Solving channel assignment problems using local search methods and simulated annealing, Independent Component Analyses, Wavelets, Neural Networks, Biosystems, and Nanoengineering IX, a part of SPIE Defense, Security, and Sensing 8058.

[41] J. Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.

[42] L.P. Wang, S. Li, S.C. Lay, W.H. Yu, C. Wan, Genetic algorithms for optimal channel assignments in mobile communications, Neural Network World 12 (6) (2002) 599–619.

[43] M. Zhu, L.P. Wang, Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms, The 2010 International Joint Conference on Neural Networks (IJCNN 2010), 2010, pp. 1–5.

[44] L.P. Wang, L. Zhou, W. Liu, FPGA segmented channel routing using genetic algorithms, IEEE Congr. Evol. Comput. (CEC 2005) 3 (2005) 2161–2165.

[45] M. Dorigo, Optimization, learning and natural algorithms, PhD thesis, Politecnico di Milano, Italy.

[46] B. Li, L.P. Wang, S. Wu, Ant colony optimization for the travelling salesman problem based on ants with memory, in: Proc. 4th International Conference on Natural Computation (ICNC 2008) 7, 2008, pp. 496–501.

[47] J. Kennedy, R. Eberhart, Particle swarm optimization, Proc. IEEE Int. Conf. Neural Networks (1995) 1942–1948.

[48] L.P. Wang, G. Si, Optimal location management in mobile computing with hybrid genetic algorithm and particle swarm optimization (pso), IEEE International Conference on Electronics, Circuits, and Systems (ICECS 2010).

[49] X. Fu, S. Lim, L.P. Wang, G. Lee, S. Ma, L. Wong, G. Xiao, Key node selection for containing infectious disease spread using particle swarm optimization, IEEE Swarm Intelligence Symposium (SIS 2009).

[50] H. Nozawa, A neural-network model as a globally coupled map and applications based on chaos, Chaos 2 (3) (1992) 377–386.

[51] L. Chen, K. Aihara, Chaotic simulated annealing by a neural network model with transient chaos, Neural Networks 8 (6) (1995) 915–930.

[52] F. Glover, Future paths for integer programming and links to artificial intelligence, Comput. Oper. Res. 13 (5) (1986) 533–549.

[53] Y. Peng, B.H. Soong, L.P. Wang, Broadcast scheduling in packet radio networks using mixed tabu-greedy algorithm, Electron. Lett. 40 (6) (2004) 375–376.

[54] L.P. Wang, S. Li, F. Tian, X. Fu, A noisy chaotic neural network for solving combinatorial optimization problems: stochastic chaotic simulated annealing, IEEE Trans. Syst. Man Cybern. Part B Cybern. 34 (5) (2004) 2119–2125.

[55] L.P. Wang, W. Liu, H. Shi, Noisy chaotic neural networks with variable thresholds for the frequency assignment problem in satellite communications, IEEE Trans. Syst. Man Cybern. Part B Cybern. 38 (2) (2008) 209–217.

[56] L.P. Wang, K. Smith, On chaotic simulated annealing, IEEE Trans. Neural Networks 9 (4) (1998) 716–718.

[57] A.H. Land, A.G. Doig, An automatic method of solving discrete programming problems, Econometrica 28 (3) (1960) 497–520.

[58] H. Shi, L.P. Wang, A mixed branch-and-bound and neural network approach for the broadcast scheduling problem, in: Proceedings of the 3rd International Conference on Hybrid Intelligent Systems (HIS 2003), 2003, 42–49.

[59] W. Siedlecki, J. Sklansky, A note on genetic algorithms for large-scale feature selection, Pattern Recogn. Lett. 10 (1989) 335–347.

[60] N. Xiong, A hybrid approach to input selection for complex processes, IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 32 (4) (2002) 532–536, http://dx.doi.org/10.1109/TSMCA.2002.804786.

[61] X. Fu, L.P. Wang, Rule extraction by genetic algorithms based on a simplified RBF neural network, in: Proceedings of the 2001 Congress on Evolutionary Computation (CEC 2001), 2001, 753–758.

[62] X. Fu, L.P. Wang, Rule extraction from an RBF classifier based on class-dependent features, in: Proceedings of the 2002 Congress on Evolutionary Computation (CEC 2002) 1, 2002, pp. 1916–1921.

[63] X. Fu, L.P. Wang, A GA-based novel rbf classifier with class-dependent features, in: Proceedings of the 2002 Congress on Evolutionary Computation (CEC 2002) 1, 2002, pp. 1890–1894.

[64] T.-C. Lin, R.-S. Liu, Y.-T. Chao, S.-Y. Chen, Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms, Gene 518 (1) (2013) 159–163. doi: 10.1016/j.gene.2012.11.046. URL http://www.sciencedirect.com/science/article/pii/S0378111912014679.

[65] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, S. Mavroudi, Yamipred: a novel evolutionary method for predicting pre-mirnas and selecting relevant features, IEEE/ACM Trans. Comput. Biol. Bioinf. 12 (5) (2015) 1183–1192, http://dx.doi.org/10.1109/TCBB.2014.2388227.

[66] P. Zhang, H. Li, H. Wang, W. Stephen, X. Zhou, Peak tree: a new tool for multiscale hierarchical representation and peak detection of mass spectrometry data, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (4) (2011) 1054–1066, http://dx.doi.org/10.1109/TCBB.2009.56.

[67] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[68] J. Zhong, J. Wang, W. Peng, Z. Zhang, M. Li, A feature selection method for prediction essential protein, Tsinghua Sci. Technol. 20 (5) (2015) 491–499, http://dx.doi.org/10.1109/TST.2015.7297748.

[69] C. Furlanello, M. Serafini, S. Merler, G. Jurman, Semisupervised learning for molecular profiling, IEEE/ACM Trans. Comput. Biol. Bioinf. 2 (2) (2005) 110–118, http://dx.doi.org/10.1109/TCBB.2005.28.

[70] K.-B. Duan, J.C. Rajapakse, H. Wang, F. Azuaje, Multiple SVM-RFE for gene selection in cancer classification with expression data, IEEE Trans. Nanobiosci. 4 (3) (2005) 228–234, http://dx.doi.org/10.1109/TNB.2005.853657.

[71] Y. Tang, Y.Q. Zhang, Z. Huang, Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis, IEEE/ACM Trans. Comput. Biol. Bioinf. 4 (3) (2007) 365–381, http://dx.doi.org/10.1109/TCBB.2007.70224.

[72] M. Yousef, S. Jung, L. Showe, M. Showe, Recursive cluster elimination (RCE) for classification and feature selection from gene expression data, BMC Bioinf. 8 (2007) 144.

[73] L.K. Luo, D.F. Huang, L.J. Ye, Q.F. Zhou, G.F. Shao, H. Peng, Improving the computational efficiency of recursive cluster elimination for gene selection, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (1) (2011) 122–129, http://dx.doi.org/10.1109/TCBB.2010.44.

[74] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, B.B. Tan, B.C. Deng, C.C. Lin, Recipe for uncovering predictive genes using support vector machines based on model population analysis, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (6) (2011) 1633–1641, http://dx.doi.org/10.1109/TCBB.2011.36.

[75] M. Hayat, M. Tahir, S.A. Khan, Prediction of protein structure classes using hybrid space of multi-profile bayes and bi-gram probability feature spaces, J. Theor. Biol. 346 (2014) 8–15, http://dx.doi.org/10.1016/j.jtbi.2013.12.015. URL.http://www.sciencedirect.com/science/article/pii/S0022519313005663.

[76] G. Bontempi, A blocking strategy to improve gene selection for classification of gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinf. 4 (2) (2007) 293–300, http://dx.doi.org/10.1109/TCBB.2007.1014.

[77] W.W.L. Wong, F.J. Burkowski, Using kernel alignment to select features of molecular descriptors in a QSAR study, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (5) (2011) 1373–1384, http://dx.doi.org/10.1109/TCBB.2011.31.

[78] I.B. Ozyurt, Automatic identification and classification of noun argument structures in biomedical literature, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (6) (2012) 1639–1648, http://dx.doi.org/10.1109/TCBB.2012.111.

[79] S. Ghorai, A. Mukherjee, S. Sengupta, P.K. Dutta, Cancer classification from gene expression data by NPPC ensemble, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (3) (2011) 659–671, http://dx.doi.org/10.1109/TCBB.2010.36.

[80] S. Fong, S. Deb, X.S. Yang, J. Li, Feature selection in life science classification: metaheuristic swarm search, IT Prof. 16 (4) (2014) 24–29, http://dx.doi.org/10.1109/MITP.2014.50.

[81] B.Y. Sun, Z.H. Zhu, J. Li, B. Linghu, Combined feature selection and cancer prognosis using support vector machine regression, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (6) (2011) 1671–1677, http://dx.doi.org/10.1109/TCBB.2010.119.

[82] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Mach. Learn. 61 (1–3) (2005) 129–150.

[83] L. Wang, J. Zhu, H. Zou, Hybrid huberized support vector machines for microarray classification and gene selection, Bioinformatics 24 (3) (2008) 412–419.

[84] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. B 67 (2005) 301–320.

[85] Z. Liu, S. Lin, M. Tan, Sparse support vector machines with LP penalty for biomarker identification, IEEE/ACM Trans. Comput. Biol. Bioinf. 7 (1) (2010) 100–107, http://dx.doi.org/10.1109/TCBB.2008.17.

[86] S. Klement, S. Anders, T. Martinetz, The support feature machine: classification with the least number of features and application to neuroimaging data, Neural Comput. 25 (6) (2013) 1548–1584, http://dx.doi.org/10.1162/NECOa00447.

[87] P. Mohapatra, S. Chakravarty, P. Dash, Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system, Swarm Evol. Comput. 28 (2016) 144–160, http://dx.doi.org/10.1016/j.swevo.2016.02.002. URL http://www.sciencedirect.com/science/article/pii/S2210650216000195.

[88] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, in: Proceedings of the 15th International Conference on Machine Learning, ICML 98, 5, 1998, pp. 242–249.

[89] S. An, W. Liu, S. Venkatesh, Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression, Pattern Recogn. 40 (8) (2007) 2154–2162.

[90] J.B. Endelman, Ridge regression and other kernels for genomic selection with r package rrBLUP, Plant Genome 4 (3) (2001) 250–255, http://dx.doi.org/10.3835/plantgenome2011.08.0024.

[91] P. Maji, S.K. Pal, Rough-fuzzy c-medoids algorithm and selection of bio-basis for amino acid sequence analysis, IEEE Trans. Knowl. Data Eng. 19 (6) (2007) 859–872, http://dx.doi.org/10.1109/TKDE.2007.190609.

[92] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.

[93] F. Fazayeli, L.P. Wang, J. Mandziuk, Feature selection based on the rough set theory and expectation-maximization clustering algorithm, Rough Sets Curr. Trends Comput. LNCS 5306 (2008) 272–282.

[94] P. Maji, P. Garai, On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance, Appl. Soft Comput. 13 (9) (2013) 3968–3980, http://dx.doi.org/10.1016/j.asoc.2012.09.006. URL http://www.sciencedirect.com/science/article/pii/S156849461200422X.

[95] U. Maulik, D. Chakraborty, Fuzzy preference based feature selection and semisupervised svm for cancer classification, IEEE Trans. Nanobiosci. 13 (2) (2014) 152–160, http://dx.doi.org/10.1109/TNB.2014.2312132.

[96] H. Pang, S.L. George, K. Hui, T. Tong, Gene selection using iterative feature elimination random forests for survival outcomes, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (5) (2012) 1422–1431, http://dx.doi.org/10.1109/TCBB.2012.63.

[97] M.Y. Wu, D.Q. Dai, Y. Shi, H. Yan, X.F. Zhang, Biomarker identification and cancer classification based on microarray data using laplace naive bayes model with mean shrinkage, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (6) (2012) 1649–1662, http://dx.doi.org/10.1109/TCBB.2012.105.

[98] V. Metsis, F. Makedon, D. Shen, H. Huang, Dna copy number selection using robust structured sparsity-inducing norms, IEEE/ACM Trans. Comput. Biol. Bioinf. 11 (1) (2014) 168–181, http://dx.doi.org/10.1109/TCBB.2013.141.

[99] M. Boareto, J. Cesar, V.B.P. Leite, N. Caticha, Supervised variational relevance learning, an analytic geometric feature selection with applications to omic datasets, IEEE/ACM Trans. Comput. Biol. Bioinf. 12 (3) (2015) 705–711, http://dx.doi.org/10.1109/TCBB.2014.2377750.

[100] M. Tan, I.W. Tsang, L. Wang, Minimax sparse logistic regression for very high-dimensional feature selection, IEEE Trans. Neural Networks Learn. Syst. 24 (10) (2013) 1609–1622, http://dx.doi.org/10.1109/TNNLS.2013.2263427.

[101] J.J.-Y. Wang, H. Bensmail, X. Gao, Feature selection and multi-kernel learning for sparse representation on a manifold, Neural Networks 51 (2014) 9–16, http://dx.doi.org/10.1016/j.neunet.2013.11.009. URL http://www.sciencedirect.com/science/article/pii/S0893608013002736.

[102] N. Garcia-Pedrajas, A. de Haro-Garcia, J. Perez-Rodriguez, A scalable approach to simultaneous evolutionary instance and feature selection, Inf. Sci. 228 (2013) 150–174, http://dx.doi.org/10.1016/j.ins.2012.10.006. URL http://www.sciencedirect.com/science/article/pii/S0020025512006718.

[103] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, A. Nowe, A survey on filter techniques for feature selection in gene expression microarray analysis, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (4) (2012) 1106–1119, http://dx.doi.org/10.1109/TCBB.2012.33.

[104] T. Marill, D. Green, On the effectiveness of receptors in recognition systems, IEEE Trans. Inf. Theory 9 (1963) 11–17.

[105] A. Whitney, A direct method of nonparametric measurement selection, IEEE Trans. Comput. 20 (1971) 1100–1103.

[106] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, E. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[107] V. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 5116–5121.

[108] M. Kerr, M. Martin, G. Churchill, Analysis of variance for gene expression microarray data, J. Comput. Biol. 7 (2000) 819–837.

[109] J.G. Thomas, J.M. Olson, S.J. Tapscott, L.P. Zhao, An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, Genome Res. 11 (2001) 1227–1236.

[110] L.J. vant Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.

[111] X. Yan, M. Deng, W.K. Fung, M. Qian, Detecting differentially expressed genes by relative entropy, J. Theory Biol. 234 (2005) 395–402.

[112] R. Steuer, J. Kurths, C.O. Daub, J. Weise, J. Selbig, The mutual information: detecting and evaluating dependencies between variables, Bioinformatics 18 (2002) S23–S240.

[113] X. Liu, A. Krishnan, A. Mondry, An entropy-based gene selection method for cancer classification using microarray data, BMC Bioinf. 6 (2005). article 76.

[114] K. Kira, L.A. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: 10th National Conference on Artificial Intelligence, 1992, pp. 129–134.

[115] I. Kononenko, Estimating attributes: Analysis and extensions of relief, in: ECML-94, 1994, pp. 171–182.

[116] R. Ruiz, J.S. Aguilar-Ruiz, J.C. Riquelme, SOAP: efficient feature selection of numeric attributes, IBERAMIA (2002) 233–242.

[117] L.-Y. Chuang, C.-H. Ke, H.-W. Chang, C.-H. Yang, A two-stage feature selection method for gene expression data, OMICS, J. Integr. Biol. 13 (2009) 127–137.

[118] F. Chu, L.P. Wang, Applications of support vector machines to cancer classification with microarray data, Int. J. Neural Syst. 15 (6) (2005) 475–484.

[119] F. Chu, L.P. Wang, Gene expression data analysis using support vector machines, Proc. Int. Joint Conf. Neural Networks 2003 (1) (2003) 2268–2271.

[120] H.C. Liu, P.C. Peng, T.C. Hsieh, T.C. Yeh, C.J. Lin, C.Y. Chen, J.Y. Hou, L.Y. Shih, D. C. Liang, Comparison of feature selection methods for cross-laboratory microarray analysis, IEEE/ACM Trans. Comput. Biol. Bioinf. 10 (3) (2013) 593–604, http://dx.doi.org/10.1109/TCBB.2013.70.

[121] N. Zhou, L.P. Wang, Effective selection of informative SNPs and classification on the HapMap genotype data, BMC Bioinf. 8 (2007) 484.

[122] N. Zhou, L.P. Wang, A modified t-test feature selection method and its application on the HapMap genotype, Genomics Proteomics Bioinf. 5 (2007) 242–249.

[123] L. Yu, Y. Han, M.E. Berens, Stable gene selection from microarray data via sample weighting, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (1) (2012) 262–272, http://dx.doi.org/10.1109/TCBB.2011.47.

[124] T. Peters, D.W. Bulger, T.H. Loi, J.Y.H. Yang, D. Ma, Two-step cross-entropy feature selection for microarrays, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (4) (2011) 1148–1151, http://dx.doi.org/10.1109/TCBB.2011.30.

[125] I. Valavanis, I. Maglogiannis, A.A. Chatziioannou, Exploring robust diagnostic signatures for cutaneous melanoma utilizing genetic and imaging data, IEEE J. Biomed. Health Inf. 19 (1) (2015) 190–198, http://dx.doi.org/10.1109/JBHI.2014.2336617.

[126] E. Gumus, Z. Gormez, O. Kursun, Multi objective SNP selection using pareto optimality, Comput. Biol. Chem. 43 (2013) 23–28, http://dx.doi.org/10.1016/j.compbiolchem.2012.12.006. URL http://www.sciencedirect.com/science/article/pii/S1476927112001156.

[127] M. Bennasar, Y. Hicks, R. Setchi, Feature selection using joint mutual information maximisation, Expert Syst. Appl. 42 (22) (2015) 8520–8532, http://dx.doi.org/10.1016/j.eswa.2015.07.007. URL http://www.sciencedirect.com/science/article/pii/S0957417415004674.

[128] P. Maji, f -information measures for efficient selection of discriminative genes from microarray data, IEEE Trans. Biomed. Eng. 56 (4) (2009) 1063–1069, http://dx.doi.org/10.1109/TBME.2008.2004502.

[129] P. Ranganarayanan, N. Thanigesan, V. Ananth, V.K. Jayaraman, V. Ramakrishnan, Identification of glucose-binding pockets in human serum albumin using support vector machine and molecular dynamics simulations, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (1) (2016) 148–157, http://dx.doi.org/10.1109/TCBB.2015.2415806.

[130] K. Leung, K. Lee, J. Wang, E.Y. Ng, H.L. Chan, S.K. Tsui, T.S. Mok, P.C.H. Tse, J.J. Sung, Data mining on dna sequences of hepatitis B virus, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (2) (2011) 428–440, http://dx.doi.org/10.1109/TCBB.2009.6.

[131] X. Xu, A. Li, M. Wang, Prediction of human disease-associated phosphorylation sites with combined feature selection approach and support vector machine, IET Syst. Biol. 9 (4) (2015) 155–163, http://dx.doi.org/10.1049/iet-syb.2014.0051.

[132] N. Zhou, L.P. Wang, Minimum redundancy feature selection from microarray gene expression data, J. Bioinf. Comput. Biol. 3 (2) (2005) 185–205.

[133] H. Mohabatkar, M.M. Beigi, A. Esmaeili, Prediction of GABAA receptor proteins using the concept of chou's pseudo-amino acid composition and support vector machine, J. Theory Biol. 281 (1) (2011) 18–23, http://dx.doi.org/10.1016/j.jtbi.2011.04.017. URL http://www.sciencedirect.com/science/article/pii/S0022519311002177.

[134] D. Wang, F. Nie, H. Huang, Feature selection via global redundancy minimization, IEEE Trans. Knowl. Data Eng. 27 (10) (2015) 2743–2755, http://dx.doi.org/10.1109/TKDE.2015.2426703.

[135] K.L. Lin, C.Y. Lin, C.D. Huang, H.M. Chang, C.Y. Yang, C.T. Lin, C.Y. Tang, D.F. Hsu, Feature selection and combination criteria for improving accuracy in protein structure prediction, IEEE Trans. Nanobiosci. 6 (2) (2007) 186–196, http://dx.doi.org/10.1109/TNB.2007.897482.

[136] C. Furlanello, S. Merler, G. Jurman, Combining feature selection and DTW for time-varying functional genomics, IEEE Trans. Signal Process. 54 (6) (2006) 2436–2443, http://dx.doi.org/10.1109/TSP.2006.873715.

[137] M. Mohammadi, H.S. Noghabi, G.A. Hodtani, H.R. Mashhadi, Robust and stable gene selection via maximum-minimum correntropy criterion, Genomics 107 (2-3) (2016) 83–87, http://dx.doi.org/10.1016/j.ygeno.2015.12.006. URL http://www.sciencedirect.com/science/article/pii/S0888754315300495.

[138] F.M. Lopes, D.C. Martins Jr., J. Barrera, R.M. Cesar Jr., A feature selection technique for inference of graphs from their known topological properties: revealing scale-free gene regulatory networks, Inf. Sci. 272 (2014) 1–15, http://dx.doi.org/10.1016/j.ins.2014.02.096. URL http://www.sciencedirect.com/science/article/pii/S0020025514002023.

[139] S. Zhang, H.S. Wong, Y. Shen, D. Xie, A new unsupervised feature ranking method for gene expression data based on consensus affinity, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (4) (2012) 1257–1263, http://dx.doi.org/10.1109/TCBB.2012.34.

[140] L.P. Wang, F. Chu, W. Xie, Accurate cancer classification using expressions of very few genes, IEEE-ACM Trans. Comput. Biol. Bioinf. 4 (2007) 40–53.

[141] X. Li, M. Yin, Multiobjective binary biogeography based optimization for feature selection using gene expression data, IEEE Trans. Nanobiosci. 12 (4) (2013) 343–353, http://dx.doi.org/10.1109/TNB.2013.2294716.

[142] Q. Wu, Y. Ye, Y. Liu, M.K. Ng, SNP selection and classification of genome-wide SNP data using stratified sampling random forests, IEEE Trans. Nanobiosci. 11 (3) (2012) 216–227, http://dx.doi.org/10.1109/TNB.2012.2214232.

[143] E. Bonilla-Huerta, A. Hernandez-Montiel, R. Morales-Caporal, M. Arjona-Lopez, Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (1) (2016) 12–26, http://dx.doi.org/10.1109/TCBB.2015.2474384.

[144] S.J. Sajjadi, X. Qian, B. Zeng, A.A. Adl, Network-based methods to identify highly discriminating subsets of biomarkers, IEEE/ACM Trans. Comput. Biol. Bioinf. 11 (6) (2014) 1029–1037, http://dx.doi.org/10.1109/TCBB.2014.2325014.

[145] B. Liu, C. Wan, L.P. Wang, An efficient semi-unsupervised gene selection method via spectral biclustering, IEEE Trans. Nano Biosci. 5 (2006) 110–114.

[146] J.X. Liu, Y. Xu, Y.L. Gao, C.H. Zheng, D. Wang, Q. Zhu, A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-Seq data, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (2) (2016) 392–398, http://dx.doi.org/10.1109/TCBB.2015.2440265.

[147] J.F.P. da Costa, H. Alonso, L. Roque, A weighted principal component analysis and its application to gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (1) (2011) 246–252, http://dx.doi.org/10.1109/TCBB.2009.61.

[148] J.X. Liu, Y. Xu, C.H. Zheng, H. Kong, Z.H. Lai, RPCA-based tumor classification using gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinf. 12 (4) (2015) 964–970, http://dx.doi.org/10.1109/TCBB.2014.2383375.

[149] S. Niijima, Y. Okuno, Laplacian linear discriminant analysis approach to unsupervised feature selection, IEEE/ACM Trans. Comput. Biol. Bioinf. 6 (4) (2009) 605–614, http://dx.doi.org/10.1109/TCBB.2007.70257.

[150] C.H. Zheng, T.Y. Ng, L. Zhang, C.K. Shiu, H.Q. Wang, Tumor classification based on non-negative matrix factorization using gene expression data, IEEE Trans. Nanobiosci. 10 (2) (2011) 86–93, http://dx.doi.org/10.1109/TNB.2011.2144998.

[151] G.R. Naik, H.T. Nguyen, Nonnegative matrix factorization for the identification of emg finger movements: evaluation using matrix analysis, IEEE J. Biomed. Health Inf. 19 (2) (2015) 478–485, http://dx.doi.org/10.1109/JBHI.2014.2326660.

[152] T. Hastie, R. Tibshirani, Efficient quadratic regularization for expression arrays, Biostatistics 5 (2004) 329–340.

[153] I. Levner, Feature selection and nearest centroid classification for protein mass spectrometry, BMC Bioinf. 6 (2005) 68.

[154] L. Li, C. Weinberg, T. Darden, L. Pedersen, Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GAKNN method, Bioinformatics 17 (2001) 1131–1142.

[155] M. Kukar, I. Kononenko, Cost-sensitive learning with neural networks, ECAI 98, in: 13th European Conference on Artificial Intelligence, 1998, 445–449.

[156] C. Wan, L.P. Wang, K.M. Ting, Introducing cost-sensitive neural networks, in: Proc. The Second International Conference on information, Communications, and Signal Processing (ICICS 99), 1999, 1B2.8.

[157] X. Fu, L.P. Wang, K.S. Chua, F. Chu, Training rbf neural networks on unbalanced data, in: Proceedings of the 9th International Conference on Neural Information Processing (ICONIP 2002), 2, 2002, 1016–1020.

[158] M. Wasikowski, X.W. Chen, Combating the small sample class imbalance problem using feature selection, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1388–1400, http://dx.doi.org/10.1109/TKDE.2009.187.

[159] S. Zhu, D. Wang, K. Yu, T. Li, Y. Gong, Feature selection for gene expression using model-based entropy, IEEE/ACM Trans. Comput. Biol. Bioinf. 7 (1) (2010) 25–36, http://dx.doi.org/10.1109/TCBB.2008.35.

[160] I.S. Oh, J.S. Lee, C.Y. Suen, Analysis of class separation and combination of class-dependent features for handwriting recognition, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 1089–1094.

[161] I.S. Oh, J.S. Lee, C.Y. Suen, Using class separation for feature analysis and combination of class-dependent features, in: Fourteenth International Conference on Pattern Recognition, vol. 1, 1998, pp. 453–455.

[162] X.J. Fu, L.P. Wang, A GA-based novel RBF classifier with class-dependent features, in: 2002 Congress on Evolutionary Computation, vol. 2, 2002, pp. 1890–1894.

[163] P.M. Baggenstoss, Class-specific features in classification, IEEE Trans. Signal Process. (2002) 3428–3432.

[164] P.M. Baggenstoss, The projection theorem and the class-specific method, IEEE Trans. Signal Process. (2003) 672–685.

[165] C.L. Liu, H. Sako, Class-specific feature polynomial classifier for pattern classification and its application to handwritten numerical recognition, Pattern Recogn. 39 (4) (2006) 669–681.

[166] L.P. Wang, N. Zhou, F. Chu, A general wrapper approach to selection of class-dependent features, IEEE Trans. Neural Networks 19 (2008) 1267–1278.

[167] Z. Zhu, Y.S. Ong, J.M. Zurada, Identification of full and partial class relevant genes, IEEE/ACM Trans. Comput. Biol. Bioinf. 7 (2) (2010) 263–277, http://dx.doi.org/10.1109/TCBB.2008.105.

[168] J.C. Rajapakse, P.A. Mundra, Multiclass gene selection using pareto-fronts, IEEE/ACM Trans. Comput. Biol. Bioinf. 10 (1) (2013) 87–97, http://dx.doi.org/10.1109/TCBB.2013.1.

[169] C. Freeman, D. Kulic, O. Basir, Feature-selected tree-based classification, IEEE Trans. Cybern. 43 (6) (2013) 1990–2004, http://dx.doi.org/10.1109/TSMCB.2012.2237394.