Artif Intell Rev
DOI 10.1007/s10462-011-9230-1

Feature selection for machine learning classification problems: a recent overview

S. B. Kotsiantis

Abstract A lot of candidate features are usually provided to a learning algorithm for producing a complete characterization of the classification task. However, it is often the case that majority of the candidate features are irrelevant or redundant to the learning task, which will deteriorate the performance of the employed learning algorithm and lead to the problem of overfitting. The learning accuracy and training speed may be significantly deteriorated by these superfluous features. So it is of fundamental importance to select the relevant and necessary features in the preprocessing step. This paper describes basic feature selection issues and current research points. Of course, a single article cannot be a complete review of all algorithms, yet we hope that the references cited will cover the major theoretical issues, guiding the researcher in interesting research directions and suggesting possible bias combinations that have yet to be explored.

Keywords Data mining · Machine learning · Pattern recognition

1 Introduction

In theory, increasing the size of the feature vector is expected to provide more discriminating power. In practice, however, excessively large feature vectors significantly slow down the learning process as well as cause the classifier to overfit the training data and compromise model generalization. Feature selection is especially important when one is handling a huge dataset with dimensions up to thousands. Feature selection has now been widely applied in many domains, such as text categorization (Forman 2003; Lee and Lee 2006), bioinformatics (Saeys et al. 2007) and astronomy (Zheng and Zhang 2008).

S. B. Kotsiantis (✉)
Educational Software Development Laboratory, Department of Mathematics, University of Patras,
P.A. Box: 1399, 26 500 Rio, Greece
e-mail: sotos@math.upatras.gr

Published online: 01 May 2011

Whereas an irrelevant feature does not affect describing the target concept in any way, a redundant feature does not add anything new to describing the target concept. Redundant features might possibly add more noise than useful information in describing the concept of interest. The main benefits of feature selection are follows: (i) reducing the measurement cost and storage requirements, (ii) coping with the degradation of the classification performance due to the finiteness of training sample sets, (iii) reducing training and utilization time and, (iv) facilitating data visualization and data understanding.

It has been shown that estimating the relevance of individual features may not be difficult; however, the real challenge is to estimate the relevance of subsets of features. This issue has been studied by a number of researchers (Peng et al. 2005; Qu et al. 2005; Yu and Liu 2004). A feature selection framework generally consists of two parts: a searching engine used to determine the promising feature subset candidates, and a criterion used to determine the best candidate.

The search for a variable subset is a NP-hard problem. Therefore, the optimal solution cannot be guaranteed to be acquired except when performing an exhaustive search in the solution space. Regarding the selection strategy, filter methods rank features or feature subsets independently of the classifier, while wrapper methods use a classifier to assess feature subsets, training one earning machine for every feature subset considered, thus, these methods are usually computationally heavy and they are conditioned to the type of classifier used (Liu and Motoda 2008). In filtering methods, a feature can be selected based upon some predefined criteria such as mutual information (Chow and Huang 2005; Sindhwani et al. 2004), independent component analysis (Plumbley and Oja 2004), class separability measure (Mao 2004), or variable ranking (Caruana and Sa 2003).

Feature construction, sometimes called feature extraction, is referred to the process of extracting new features by transforming the original sample data set representation, in order to have the problem represented in a more discriminative (informative) space that makes the classification task more efficient. Principal component analysis (Malhi and Gao 2004) and other transformation-based feature reduction methods are not discussed in this paper because they do not select the features from the original feature set. These methods transform the feature set into a lower-dimensional feature vector by combining several features.

Another category of approaches called feature weighting approaches, is not always considered in the classical classification of feature selection methods. In the implementation process of these methods, actual feature selection is substituted by a feature weighing procedure able to weight the relevance of the features.

Supervised feature selection determines relevant features by their relations with the corresponding class labels and discards irrelevant and redundant features. We have limited our references to recent refereed journals, published books and conferences. A previous review of feature selection can be found in Guyon and Elisseeff (2003). The reader should be cautioned that a single article cannot be a comprehensive review of all feature selection algorithms. Instead, our goal has been to provide a representative sample of existing lines of research in feature selection. In each of our listed areas, there are many other papers that more comprehensively detail relevant work.

Our next section covers basic issues of feature selection for classification problems. In Sect. 3, we are referred to search strategies. Section 4 deals with evaluation criteria. In Sect. 5, semi-supervised feature selection techniques are presented. Finally, the last section concludes this work.

2 Basic issues of feature selection

In most real-world datasets, not all of the attributes contribute to the definition or determination of class labels. In theory, increasing the size of the feature vector is expected to provide more discriminating power. In practice, however, excessively large feature vectors significantly slow down the learning process as well as cause the classifier to overfit the training data and compromise model generalization (Hall and Holmes 2003).

Moreover, the cost for measuring a feature is a critical issue to be considered while selecting a subset. In case of medical diagnosis the features may be observable symptoms or diagnostic tests. Each clinical test is associated with its own diagnostic value, cost and risk. The challenge is in selecting the subset of features with minimum risk, least cost yet which is significantly important in the determining its class/pattern.

Generally, features are characterized as: (i) Relevant: features which have an influence on the output and their role can not be assumed by the rest, (ii) Irrelevant: features not having any influence on the output, (iii) Redundant: a feature can take the role of another.

The goal of feature selection is to find the optimal subset consisting of m features chosen from the total n features. One critical problem for many feature selection methods is that an exhaustive search strategy has to be applied to seek the best subset among all the possible $\binom{n}{m}$ feature subsets, which usually results in a considerably high computational complexity. The alternative suboptimal feature selection methods provide more practical solutions in terms of computational complexity but they cannot promise that the finally selected feature subset is globally optimal.

The relevance of the features can be evaluated either individually (univariate approaches), or in a multivariate manner. Univariate approaches are simple and fast, therefore, appealing. However, possible correlation and dependencies between the features are not considered. Therefore, multivariate search techniques may be helpful. Several limitations restrict the use of multivariate searches. First, they are prone to overtraining, especially in $p \gg n$ (many features and few samples) settings. Secondly, they can be computationally too expensive when dealing with a large feature space.

Even when feature ranking is not optimal, it may be preferable to other variable subset selection methods because of its computational and statistical scalability: Computationally, it is efficient since it requires only the computation of n scores and sorting the scores. One common criticism of feature ranking is that it leads to the selection of a redundant subset. The same performance could possibly be achieved with a smaller subset of complementary features. Moreover, a feature that is completely useless by itself can provide a significant performance improvement when taken with others. Two features that are useless by themselves can be useful together. Probably the simplest solution with filtering techniques consists in fixing in advance the feature subset size k. Feature ranking can also rank the features by a metric and eliminates all features that do not achieve an adequate score. Stoppiglia et al. (2003) proposed to append to the set of candidate features a "probe" feature, which is a random variable. All features that are ranked below the probe feature should be discarded.

Advantages of filter methods are that they are fast and easy to interpret. The characteristics of filter methods are as follows: (i) Features are considered independently, (ii) Redundant features may be included, (iii) Some features which as a group have strong discriminatory power but are weak as individual features will be ignored, and (iv) The filtering procedure is independent of the classifying method.

The characteristics of wrapper methods are listed below: (i) Computationally expensive for each feature subset considered, since the classifier is built and evaluated, (ii) As exhaustive

searching is impossible, only greedy search is applied. The advantage of greedy search is simple and quickly to find solutions, but its disadvantage is not optimal, and susceptible to false starts, (iii) It is often easy to overfit in these methods.

Because of the difficulty associated with estimating the relevance of subsets of features, Al-Ani (2009) adopts the wrapper approach and focuses on dependency between feature pairs as a mean to guide the search.

More filters (e.g. those based on mutual information criteria) provide a generic selection of features, not tuned for/by a given learning machine. Another compelling justification is that filtering can be used as a preprocessing step to reduce space dimensionality and overcome overfitting. In that respect, it seems reasonable to use a hybrid technique that combines filtering and wrapper methods (Zhu et al. 2007).

Another type of feature subset selection is identified as embedded methods. In this case, the feature selection process is done inside the induction algorithm itself, i.e. attempting to jointly or simultaneously train both a classifier and a feature subset. They often optimize an objective function that jointly rewards the accuracy of classification and penalizes the use of more features.

As we have already mentioned, a feature selection framework generally consists of two parts: a searching engine used to determine the promising feature subset candidates, and a criterion used to determine the best candidate.

3 Search strategies

The exhaustive feature subset set is too costly and practically prohibitive, even for a medium-sized feature set size. Other methods based on heuristic or optimization techniques; attempt to reduce computational complexity by compromising performance.

A good search algorithm should provide: (1) good global search capability that allows for the exploration of new regions of the solution space without getting stuck in local minima, (2) rapid convergence to a near optimal solution, (3) good local search ability, and (4) high computational efficiency.

3.1 Greedy and heuristic search

The most popular feature selection methods apply forward or backward sequential schemes, which always provide a sub-optimal solution. Forward strategies usually provide a nested rank of variables, with the drawback of conditioning the m selected features given the previous m-1 selected. The backward strategy is analogous to the forward one, but starting from the whole set of variables and discarding one at the time to get to the subset of m desired features.

Sequential Forward Floating Selection (SFFS) improves the forward sequential method by introducing backward steps after each forward step as long as the objective criteria increase. In detail, the SFFS algorithm begins the search with an empty feature set and uses the basic SFS algorithm to add one feature at a time to the selected feature subset (Somol et al. 2004). Every time a new feature is added to the current feature set, the algorithm attempts to backtrack by using the SBS algorithm to remove one feature at a time to locate a better subset. The algorithm stops when the size of the current feature set is larger than the number of features d we want. This is necessary to allow backtracking. The SBFS method starts with all input features D, removes one feature at a time, and conditionally adds a feature to the resultant subset as long as a better subset can be located.

Liu and Zheng (2006) present another feature selection method named filtered and supported sequential forward search (FS_SFS) in the context of support vector machines (SVM). In comparison with conventional wrapper methods that employ the SFS strategy, FS_SFS has two important properties to reduce the time of computation. First, it dynamically maintains a subset of samples for the training of SVM. Because not all the available samples participate in the training process, the computational cost to obtain a single SVM classifier is decreased. Secondly, a new criterion, which takes into consideration both the discriminant ability of individual features and the correlation between them, is proposed to effectively filter out nonessential features.

In Tang and Mao (2007), a criterion avoids feature-type transformation through carefully decomposing the feature space along values of nominal features in the mixed feature subset and measuring class separability based on continuous features in each subspace generated and then combining these measures to produce an overall evaluation. The search algorithm, named mixed forward selection (MFS), is different from traditional search algorithms because it considers the feature-type in both subsets generation and comparison.

An improved forward floating selection (IFFS) algorithm for selecting a subset of features is presented by Nakariyakul and Casasent (2009). The proposed algorithm improves the state-of-the-art sequential forward floating selection algorithm. The improvement is to add an additional search step called "replacing the weak feature" to check whether removing any feature in the currently selected feature subset and adding a new one at each sequential step can improve the current feature subset.

Wang et al. (2009) propose a hybrid feature selection approach using tabu search and probabilistic neural networks. This tabu search algorithm uses a long-term memory to avoid the necessity of the delicate tuning of the memory length and to decrease the risk of generating a cycle that traps the search in local optimal solutions.

Yusta (2009) presents three metaheuristic strategies to solve the feature selection problem, specifically, GRASP, Tabu Search and the Memetic Algorithm. These three strategies are compared with a Genetic Algorithm and with other typical feature selection methods, such as Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SBFS). The results show that, in general, GRASP and Tabu Search obtain significantly better results than the other methods.

3.2 Optimization based search

Yang and Olafsson (2006) analyze an optimization-based feature selection method. First, they show how the use of random instance sampling can considerably reduce the computational time of the algorithm, sometimes by an order of magnitude, and thus improve its scalability. Second, given that such sampling is found to be very effective, they develop a new adaptive version of the algorithm that automatically adjusts the instance sampling rate.

Angelis et al. (2006) show how the Feature Selection problem can be formulated as a subgraph selection problem derived from the lightest k-subgraph problem, and solved as an Integer Program. Wang and Huang (2009) proposed another criterion for single/multiple objective evolutionary feature selection.

Since the feature selection problem can be considered as a combination optimization problem, researchers have used: Genetic algorithms, Ant Colony Optimization, Rough set methods and Particle Swarm Optimization.

3.2.1 Genetic algorithms

Another search procedure is based on the Genetic Algorithm (GA), which is a combinatorial search technique based on both random and probabilistic measures. Subsets of features are evaluated using a fitness function and then combined via cross-over and mutation operators to produce the next generation of subsets.

In Oh et al. (2004), hybrid GAs are proposed that include local search operators to improve the fine-tuning capabilities of simple GAs. The local search operators allow adding (removing) the most (least) significant feature to individuals in the GA population.

The conventional feature selection method with genetic algorithm has difficulty for huge-scale feature selection. Hong and Cho (2006) modify the representation of chromosome to be suitable for huge-scale feature selection and adopt speciation to enhance the performance of feature selection by obtaining diverse solutions.

In the study of Huang et al. (2007), a hybrid genetic algorithm is adopted to find a subset of features that are most relevant to the classification task. Two stages of optimization are involved. The outer optimization stage completes the global search for the best subset of features in a wrapper way, in which the mutual information between the predictive labels of a trained classifier and the true classes serves as the fitness function for the genetic algorithm. The inner optimization performs the local search in a filter manner, in which an improved estimation of the conditional mutual information acts as an independent measure for feature ranking taking account of not only the relevance of the candidate feature to the output classes but also the redundancy to the already-selected features.

Huang et al. (2008) focus on enhancing the effectiveness of filter feature selection models from two aspects. First, feature-searching engine is modified based on optimization theory. Second, a point injection strategy is designed to improve the regularization capability of feature selection. They apply these strategies to modify two typical filter models—SFS-based and GA-based approaches.

More recently, Gheyas and Smith (2010) present a hybrid algorithm (SAGA), named after two major underlying search algorithms (SA and GA), for selecting optimal feature subsets efficiently. This algorithm is based on a simulated annealing, a genetic algorithm, a generalized neural networks and a greedy search algorithm.

3.2.2 Ant colony optimization

In order to solve an optimization problem, a number of artificial ants are used to iteratively construct solutions. In each iteration, an ant would deposit a certain amount of pheromone proportional to the quality of the solution. At each step, every ant computes a set of feasible expansions to its current partial solution and selects one of these depending on two factors: local heuristics and prior knowledge. It is worth mentioning that Ant colony optimization (ACO) makes probabilistic decision in terms of the artificial pheromone trails and the local heuristic information. This allows ACO to explore larger number of solutions than greedy heuristics. Another characteristic of the ACO algorithm is the pheromone trail evaporation, which is a process that leads to decreasing the pheromone trail intensity over time (Parpinelli et al. 2002).

In Ani (2005) an Ant Colony Optimization (ACO) approach was presented for feature selection problems. The author calculates a term called "updated selection measure (USM)" which is used for selecting features, a function of the pheromone trail and the so called "local importance" which has replaced the heuristic function. Sivagaminathan and Ramakrishnan (2007) also present a hybrid method based on ant colony optimization and artificial neural

networks (ANNs) to address feature selection. In Kanan and Faez (2008), a modified ACO-based feature selection algorithm has been introduced, too. The classifier performance and the length of the selected feature vector are adopted as heuristic information for ACO.

3.2.3 Rough set methods

Rough set theory has been introduced by Pawlak (1991) to deal with imprecise or vague concepts. Swiniarski and Skowron (2003) present applications of rough set methods for feature selection.

Hu et al. (2008a) generalize Pawlak's rough set model into d neighborhood rough set model and k-nearest-neighbor rough set model, where the objects with numerical attributes are granulated with d neighborhood relations or k-nearest-neighbor relations, while objects with categorical features are granulated with equivalence relations. Then the induced information granules are used to approximate the decision with lower and upper approximations. The authors compute the lower approximations of decision to measure the significance of attributes. Based on the proposed models, they give the definition of significance of mixed features and construct a greedy attribute reduction algorithm.

Chen et al. (2008) propose a bit-based feature selection method to find the smallest feature set to represent the indexes of a given dataset. That approach originates from the bitmap indexing and rough set techniques. It consists of two-phases. In the first phase, the given dataset is transformed into a bitmap indexing matrix with some additional data information. In the second phase, a set of relevant and enough features are selected and used to represent the classification indexes of the given dataset. After the relevant and enough features are selected, they can be judged by the domain expertise and the final feature set of the given dataset is thus proposed.

Hu et al. (2008b) show some metrics to compute neighborhoods of samples in general metric spaces, and then they introduce the neighborhood rough set model and discuss the properties of neighborhood decision tables. Based on the proposed model, the dependency between heterogeneous features and decision is defined for constructing measures of attribute significance for heterogeneous data.

Feature selection is a key issue in the research on rough set theory. However, in handling large-scale data, many current feature selection methods based on rough set theory are incapable. In Jiao et al. (2010), two novel feature selection methods are put forward based on decomposition and composition principles. The idea of decomposition and composition is to break a complex table down into a master-table and several sub-tables that are simpler, more manageable and more solvable by using existing induction methods, then joining them together in order to solve the original table.

More recently, Chen et al. (2010) propose a rough set approach to feature selection based on ACO, which adopts mutual information based feature significance as heuristic information. Salamó and López-Sánchez (2011) addresses the feature selection from the filter perspective, presenting three different selection strategies and several measures for estimating attribute relevance based on Rough Set Theory.

3.2.4 Particle swarm optimization

Particle swarm optimization (PSO) is a population-based stochastic optimization technique, and was developed by Kennedy and Eberhart (1995). PSO simulates the social behavior of organisms, such as bird flocking and fish schooling, to describe an automatically evolving

system. In PSO, each single candidate solution is "an individual bird of the flock", that is, a particle in the search space. Each particle makes use of its individual memory and knowledge gained by the swarm as a whole to find the best solution.

Wang et al. (2007) propose a feature selection strategy based on rough sets and particle swarm optimization (PSO). Rough sets have been used as a feature selection method with much success, but current hill-climbing rough set approaches to feature selection are inadequate at finding optimal reductions as no perfect heuristic can guarantee optimality. On the other hand, complete searches are not feasible for even medium-sized datasets. So, stochastic approaches provide a promising feature selection mechanism. Like Genetic Algorithms, PSO is an evolutionary computation technique, in which each potential solution is seen as a particle with a certain velocity flying through the problem space. Compared with GAs, PSO does not need complex operators such as crossover and mutation, it requires only primitive and simple mathematical operators, and is computationally inexpensive in terms of both memory and runtime.

More recently, Bae et al. (2010) propose an evolutionary algorithm called Intelligent Dynamic Swarm (IDS) that is a modified Particle Swarm Optimization.

4 Evaluation criteria

Class separability is a classical criterion of filtering available in the literature. It involves calculating the normalized distance between classes and then eliminating the features that yield low separability values. The criterion is computationally simple and thus satisfies the first requirement. However, it has a major drawback, i.e., the criterion implicitly assumes the features to be orthogonal and overlooks the correlation between them. Consequently, those correlated features that individually separate the classes well but collectively provided redundant information might be retained (Reunanen 2003).

Quite commonly, the researchers focus on the design of performance measures to determine the relevance between features and decision. Distance, consistency, correlation, mutual information, and dependent criteria are usually used (Guyon and Elisseeff 2003).

4.1 Distance and consistency measures

Piramuthu (2004) evaluates several inter-class as well as probabilistic distance-based feature selection methods as to their effectiveness in preprocessing input data for inducing decision trees. Results from this study show that inter-class distance measures result in better performance compared to probabilistic measures, in general.

Among feature selection techniques, the Relief algorithm is one of the most common due to its simplicity and effectiveness (Sun 2007). The performance of the Relief algorithm, however, could be dramatically affected by the consistency of the data patterns. For instance, Relief-F could become less accurate in the presence of noise. The accuracy would decrease further if an outlier sample was included in the dataset. Therefore, it is very important to select the samples to be included in the dataset carefully. Saethang et al. (2009) present an effort to improve the effectiveness of Relief algorithm by filtering samples before selecting features.

Liang et al. (2008) propose a feature selection algorithm based on a distance discriminant (FSDD), which not only solves the problem of the high computational costs but also overcomes the drawbacks of the suboptimal methods. This method is able to find the optimal feature subset without exhaustive search or Branch and Bound algorithm. The most difficult

problem for optimal feature selection, the search problem, is converted into a feature ranking problem following rigorous theoretical proof such that the computational complexity can be greatly reduced.

Recently, Hu et al. (2010) propose a concept of neighborhood margin and neighborhood soft margin to measure the minimal distance between different classes. They use the criterion of neighborhood soft margin to evaluate the quality of candidate features and construct a forward greedy algorithm for feature selection.

Lashkia and Anthony (2004) first prove that the notion of relevance of attributes is directly related to the consistency of attributes, and show how relevant, irredundant attributes can be selected. They then compare different relevant attribute selection algorithms, and show the superiority of algorithms that select irredundant attributes over those that select relevant attributes. They also show that searching for an "optimal" subset of attributes, which is considered to be the main purpose of attribute selection, is not the best way to improve the accuracy of classifiers. Employing sets of relevant, irredundant attributes improves classification accuracy in many more cases. Finally, they propose a method for selecting relevant examples, which is based on filtering the so-called pattern frequency domain.

Kahramanli et al. (2011) propose a new decision relative discernibility function (DF)-based FS approach. But in comparison with existing ones of this kind, it looks at every dataset with binary-encoded values of features as a logic truth table for the given decision function and process it in accordance to the peculiarities of FS.

4.2 Correlation measures and information theoretic ranking criteria

There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory.

Perfectly correlated features are truly redundant in the sense that no additional information is gained by adding them. In Bhavani et al. (2008), a filter-based feature selection method based on correlation fractal dimension (CFD) discrimination measure is proposed. They demonstrate through experimentation on data sets of various sizes that fractal dimension-based algorithms cannot be applied routinely to higher dimensional data sets as the calculation of fractal dimension is inherently sensitive to parameters like range of scales and the size of the data sets. Based on the empirical analysis, they propose a feature selection technique using CFD that avoids the above issues.

Among the mutual information (MI) estimators that can be found in the literature, the estimator proposed by Kraskov et al. (2004), based on K-nearest neighbors distances, has been widely employed for its data efficiency (working with low number of samples) and its high estimation performances. However, this estimator was originally proposed for regression problems. When it is used for classification tasks, it requires to code the classes with numerical values. Gomez-Verdejo et al. (2009) introduce a new MI estimator derived from Kraskov's one, dedicated to classification problems.

Some missing values of features arisen from various aspects may exist within the data sets. They will frustrate the performance of the selection and learning algorithms. Thus, it would be better if they were pre-processed before being fed into the feature selectors. For the continuous features, their information entropies can be calculated by integral form, Parzan window and Gaussian kernel-based techniques (Huang and Chow 2005). Hild et al. (2006) also estimated the quadratic entropy using Parzen windows and Gaussian kernels, instead of estimating Shannon's entropy, thus reducing the computational complexity.

Abe and Kudo (2006) propose another classifier-independent feature selection method on the basis of the estimation of Bayes discrimination boundary.

Combining the mutual information criterion with a forward feature selection strategy offers a good trade-off between optimality of the selected feature subset and computation time. However, it requires to set the parameter(s) of the mutual information estimator and to determine when to halt the forward procedure. These two choices are difficult to make because, as the dimensionality of the subset increases, the estimation of the mutual information becomes less and less reliable. Francois et al. (2007) propose to use resampling methods, a K-fold cross-validation and the permutation test, to address both issues. The resampling methods bring information about the variance of the estimator, information which can then be used to automatically set the parameter and to calculate a threshold to stop the forward procedure.

Guo et al. (2008) show that there is a general framework based on the criterion of mutual information (MI) that can provide a realistic solution to the problem of feature selection for high-dimensional data. They give a theoretical argument showing that the MI of multi-dimensional data can be broken down into several one-dimensional components, which makes numerical evaluation much easier and more accurate. Although there is a direct way of selecting features by jointly maximising MI, this suffers from combinatorial explosion. Hence, Guo et al. (2008) propose a fast feature-selection scheme based on a 'greedy' optimisation strategy.

In Bonev et al. (2008), entropic spanning graphs are used to estimate the mutual information between high-dimensional set of features and the classes. In this method, entropies are estimated directly from data samples. In this approach, the complexity does not depend on the number of dimensions but on the number of samples. However, a greedy forward feature selection algorithm is used, which adds features one at a time.

Li and Yang (2008) employed bootstrapping sampling technique to obtain multiple feature subsets by virtue of mutual information criterion, and then optimally integrated them into one using SVM.

In Jin-Jie et al. (2008), other two information-theoretic measures for feature ranking are presented: one is an improved formula to estimate the conditional mutual information between the candidate feature and the target class; the other is a mutual information (MI) based constructive criterion that is able to capture both irrelevant and redundant input features under arbitrary distributions of information of features.

Meyer et al. (2008) present a filter approach for effective feature selection in microarray data characterized by a large number of input variables and a few samples. The approach is based on the use of a information-theoretic selection, the Double Input Symmetrical Relevance (DISR), which relies on a measure of variable complementarity. This measure evaluates the additional information that a set of variables provides about the output with respect to the sum of each single variable contribution.

Liu et al. (2009b) propose a general scheme of boosting feature selection method using information metric. The primary characteristic of their method is that it exploits weight of data to select salient features. Furthermore, the weight of data will be dynamically changed after each candidate feature has been selected. Thus, the information criteria used in feature selector can exactly represent the relevant degree between features and the class labels.

Recently, a filter method of feature selection based on mutual information, called normalized mutual information feature selection (NMIFS), is presented in Estévez (2009). In Sotoca and Pla (2010) a supervised feature selection approach is presented, which is based on metric applied on continuous and discrete data representations. This method builds a dissimilarity space using information theoretic measures, in particular conditional mutual information between features with respect to a relevant variable that represents the class labels. Applying a hierarchical clustering, the algorithm searches for a compression of the information contained in the original set of features.

4.3 Fuzzy evaluation measures

Hu et al. (2006a) extended Shannon's entropy to measure the information quantity in a set of fuzzy sets and applied the proposed measure to calculate the uncertainty in fuzzy approximation spaces and used it to reduce heterogeneous data (Hu et al. 2006b), where numerical attributes induce fuzzy relations and symbolic features generate crisp relations, then the generalized information entropy is used to compute the information quantity introduced by the corresponding feature or feature subset.

The feature saliency measure proposed in Verikas et al. (2008) is based on two factors, namely, the fuzzy derivative of the predictor output with respect to the feature and the similarity between the feature and the feature set. By using the concept of fuzzy derivative instead of the ordinary derivative they model the vagueness that occurs in estimating the predictor output sensitivity. To assess the similarity of features, they resorted to the so-called maximal information compression index. To make a decision about inclusion of a candidate feature into a feature set, the paired t test comparing the saliency of the candidate and the noise feature is used.

Sánchez et al. (2008) propose an extended definition of the mutual information between two fuzzified continuous variables. They also introduce a numerical algorithm for estimating the mutual information from a sample of vague data. They showed that this estimation can be included in a feature selection algorithm, and also that, in combination with a genetic optimization, the same definition can be used to obtain the most informative fuzzy partition for the data.

An efficient filter feature selection (FS) method is proposed in Moustakidis and Theocharis (2010), the SVM-FuzCoC approach. To assess the quality of features, the authors introduce a local fuzzy evaluation measure with respect to patterns that embraces fuzzy membership degrees of every pattern in their classes. Accordingly, the above measure reveals the adequacy of data coverage provided by each feature. The required membership grades are determined via a fuzzy output kernel-based support vector machine, applied on single features. Based on a fuzzy complementary criterion (FuzCoC), the FS procedure iteratively selects features with maximum additional contribution in regard to the information content provided by previously selected features. This search strategy leads to small subsets of powerful and complementary features, alleviating the feature redundancy problem.

In Luukka (2011), feature selection method based on fuzzy entropy measures is introduced and it is tested together with similarity classifier.

4.4 Dependent criteria

For supervised learning, the primary goal of classification is to maximize predictive accuracy, therefore, predictive accuracy is generally accepted and widely used as the primary measure by researchers and practitioners.

Maldonado and Weber (2009) introduce a wrapper algorithm for feature selection, using Support Vector Machines with kernel functions. Their method is based on a sequential backward selection, using the number of errors in a validation subset as the measure to decide which feature to remove in each iteration.

By using the learning machine as a black box, wrappers are remarkably universal and simple. But embedded methods that incorporate feature selection as part of the training process may be more efficient in several respects: they make better use of the available data by not needing to split the training data into a training and validation set.

Commonly, when a feature selection algorithm is applied, a single feature subset is selected for all the classes, but this subset could be inadequate for some classes. Class-specific feature selection allows selecting a possible different feature subset for each class. In this kind of algorithms different approaches have been proposed (Fu and Wang 2005) which are strongly related to the use of a particular classifier.

4.4.1 Hybrid technique

Uncu and Turksen (2007) propose a feature selection algorithm that avoids the problem of over-fitting by first filtering the potential significant features or feature subset combinations and then identifying the best input variable combination by means of a wrapper.

Ng et al. (2008) propose a hybrid filter–wrapper-type feature subset selection methodology using a localized generalization error model. The localized generalization error model for a radial basis function neural network bounds from above the generalization error for unseen samples located within a neighborhood of the training samples. Iteratively, the feature making the smallest contribution to the generalization error bound is removed.

Liu et al. (2009a) developed a wrapper-based optimized SVM model for demand forecasting. Wrappers based on the genetic algorithm are employed to analyze the sales data of a product.

Both filter and wrapper approaches to feature selection are also combined in Bacauskiene et al. (2009). In the first phase, definitely redundant features are eliminated based on the paired *t* test. The test compares the saliency of the candidate and the noise features. In the second phase, the genetic search is employed. The search integrates the steps of training, aggregation of committee members, selection of hyper-parameters, and selection of salient features into the same learning process.

The main conclusion of Lutu and Engelbrecht (2010) is that, decision rule-based feature selection enables one to incorporate domain-specific definitions of feature relevance into the feature selection process. When a feature subset search algorithm uses decision rules to guide the search, the algorithm makes better decisions compared to when mathematical functions are used. This leads to the selection of features that provide a high level of predictive classification performance.

Unler et al. (2010) presents a hybrid filter–wrapper feature subset selection algorithm based on particle swarm optimization (PSO) for support vector machine (SVM) classification. The filter model is based on the mutual information and is a composite measure of feature relevance and redundancy with respect to the feature subset selected. The wrapper model is a modified discrete PSO algorithm.

The approach proposed by Peng et al. (2010) is featured by (1) adding a pre-selection step to improve the effectiveness in searching the feature subsets with improved classification performances and (2) using Receiver Operating Characteristics (ROC) curves to characterize the performance of individual features and feature subsets in the classification.

Lee and Leu (2011) propose a novel hybrid method for feature selection in microarray data analysis. This method first uses a genetic algorithm with dynamic parameter setting (GADP) to generate a number of subsets of genes and to rank the genes according to their occurrence frequencies in the gene subsets. Then, this method uses the x^2 test for homogeneity to select a proper number of the top-ranked genes for data analysis.

A hybrid feature selection mechanism is also proposed in Hsu et al. (2011). The idea is to utilize the efficiency of filters and the accuracy of wrappers. A three-step procedure including preliminary screening, combination, and fine tuning, was designed. Preliminary

screening and combination can quickly remove most irrelevant features. Fine tuning then further examines the combined feature set.

Bermejo et al. (2011) propose an algorithm that iteratively alternates between filter ranking construction and wrapper feature subset selection (FSS). Thus, the FSS only uses the first block of ranked attributes and the ranking method uses the current selected subset in order to build a new ranking where this knowledge is considered. The algorithm terminates when no new attribute is selected in the last call to the FSS algorithm. The main advantage of this approach is that only a few blocks of variables are analyzed, and so the number of wrapper evaluations decreases drastically.

Xie and Wang (2011) propose a hybrid feature selection method, named improved F-score and Sequential Forward Search (IFSFS). They improved the original F-score from measuring the discrimination of two sets of real numbers to measuring the discrimination between more than two sets of real numbers. The improved F-score and Sequential Forward Search (SFS) are combined to find the optimal feature subset in the process of feature selection, where, the improved F-score is an evaluation criterion of filter method, and SFS is an evaluation system of wrapper method. The best parameters of kernel function of SVM are found out by grid search technique.

4.4.2 Embedded methods

Lai et al. (2006) apply a multivariate search technique on a subspace randomly selected from the original feature space. In order to take into account all the measured features of the dataset, the procedure is repeated many times. As a result several feature subsets are selected. These are combined into a final list of selected features, by ordering the features based on their relevance derived from their accuracy in the individuals runs. The final classifier can then be trained by using the final list of features.

Feature selection and feature weighting are useful techniques for improving the classification accuracy of K-nearest-neighbor (K-NN) rule. Tahir et al. (2007) propose an approach for simultaneous feature selection and feature weighting of K-NN rule based on Tabu Search (TS) heuristic.

ElAlami (2009) describes another feature subset selection algorithm, which utilizes a genetic algorithm (GA) to optimize the output nodes of trained artificial neural network (ANN). The algorithm does not depend on the ANN training algorithms or modify the training results. The two groups of weights between input-hidden and hidden-output layers are extracted after training the ANN on a given database. The general formula for each output node (class) of ANN is then generated. This formula depends only on input features because the two groups of weights are constant. This dependency is represented by a non-linear exponential function. The GA is involved to find the optimal relevant features, which maximize the output function for each class. The dominant features in all classes are the features subset to be selected from the input feature group.

Li and Lu (2009) propose a feature selection criterion based on low-loss nearest neighbor classification and a feature selection algorithm that optimizes the margin of nearest neighbor classification through minimizing its loss function. At the same time, theoretical analysis based on energy-based model is presented.

Nguyen and Torre (2010) propose a convex energy-based framework to jointly perform feature selection and SVM parameter learning for linear and non-linear kernels.

When constructing a Bayesian network classifier from data, the more or less redundant features included in a dataset may bias the classifier and as a consequence may result in

relatively poor classification accuracy. Drugan and Wiering (2010) study the problem of selecting appropriate subsets of features for such classifiers.

Kabir et al. (2010) present a feature selection algorithm based on the wrapper approach using neural networks (NNs). The vital aspect of this algorithm is the automatic determination of NN architectures during the FS process. The algorithm uses a constructive approach involving correlation information in selecting features and determining NN architectures.

4.4.3 Class-specific feature selection

In Fu and Wang (2005) a different feature subset is selected for each class, and a RBF (Radial Based Function) classifier is proposed. This work is based on RBF neural networks, which have a set of hidden units, each one used for identifying one class, therefore, a subset of hidden units can be used for discriminating a class from the others. For identifying a feature subset for each class, a genetic algorithm (GA) determines a feature mask for the hidden unit subset associated to each class.

In Wang et al. (2008), a FS algorithm has been proposed which first converts a C class classification problem in C two-class classification problems. It means the examples in a training set are divided into two classes (say, C1 and C2). For finding feature subset of each binary classification problem, the FS algorithm then integrates features in Sequential Forward Selection (SFS) manner for training a support vector machine.

Pineda-Bautista et al. (2011) propose a general framework for using any traditional feature selector for doing class-specific feature selection, which allows using any classifier. The framework proposes to use the one-against-all class binarization method for transforming a c-class problem into c binary problems, one for each class, where the in-stances of a class are used as positive examples, and all other instances as negatives. For doing class-specific feature selection, traditional feature selectors are applied over these binary problems, thus, the feature subset selected for each binary problem is assigned to the class from which this problem was constructed. In order to classify new instances the framework proposes to use a classifier ensemble, where, for each class, a classifier is trained using the whole training set, but using the feature subset assigned to the class.

5 Semi-supervised feature selection methods

When given sufficient labeled data, the supervised feature selection methods usually outperform the unsupervised feature selection methods, since the former ones can effectively evaluate the correlation between the features and the class labels, while the latter ones only evaluate feature relevance by the capability of preserving some properties of the given data such as the variance.

Lee et al. (2006) propose an enhanced filter method that exploits features from two information-based filtering steps: supervised and unsupervised. By combining the features in these steps the authors attempt to reduce biases caused by misleading causal relations induced in the supervised selection procedure.

Usually only a few labeled data are obtained because obtaining class labels is expensive but many unlabeled data can be easily obtained. For this case, directly using the existing supervised feature selection algorithms might fail, since the data distribution may not be accurately estimated only by using a few labeled data. So, currently, an important strategy on this direction, called semi-supervised feature selection, is to simultaneously use both labeled and unlabeled data for feature selection (Kumar and Kummamuru 2008), which incorporates

the popular semi-supervised learning technique into the existing feature selection models. However, like the existing supervised feature selection methods, the supervision information used in semi-supervised feature selection is still class labels.

In fact, in contrast to class labels, other forms of supervision information such as pairwise constraints can be more easily obtained because pairwise constraint can be easily generated using class labels, the so-called pairwise constraints specifies whether a pair of data samples belong to the same class (must-link constraints) or different classes (cannot-link constraints) (Zhang et al. 2008; Yang and Song 2010)

Zhao et al. (2008) define another score which uses both the pairwise constraints defined by the user and the unlabeled nearest neighbors of the samples. However, this score considers the neighbors of each sample without explicitly taking into account its local density property. Recent constraint scores developed for semi-supervised feature selection purpose have been presented in Kalakech et al. (2011).

One major disadvantage of the Constraint Score is that its performance is dependent on a good selection on the composition and cardinality of constraint set, which is very challenging in practice. Sun and Zhang (2010) try to address the problem by importing Bagging into Constraint Score and a method called Bagging Constraint Score (BCS). Instead of seeking one appropriate constraint set for single Constraint Score, in BCS they perform multiple Constraint Score, each of which uses a bootstrapped subset of original given constraint set.

Liu et al. (2009a) propose a feature selection algorithm based on dynamic mutual information, which is only estimated on unlabeled instances. In this method, mutual information of each candidate feature is re-calculated on unlabeled instances, rather than the whole sampling space. The advantage of this approach is that it can exactly measure the relevance between candidate feature and the class labels by following the selection procedure.

6 Conclusion

Given an input set of features, dimensionality reduction can be achieved in two different ways. The first is to select the hopefully best subset of features of the input feature set. This process is termed feature selection (FS). The second approach that creates new features based on transformation from the original features to a lower dimensional space is termed feature extraction (FE). This transformation may be a linear or nonlinear combination of the original features. The choice between FS and FE depends on the application domain and the specific training data which are available. FS leads to savings in measurements cost since some of the features are discarded and the selected features retain their original physical interpretation. In addition, the retained features may be important for understanding the physical process that generates the patterns. On the other hand, transformed features generated by FE may provide a better discriminative ability than the best subset of given features, but these new features may not have a clear physical meaning.

About the feature search procedure, the number of subsets that can be extracted from the initial set is exponential in the number of initial features. In most cases, this results in the impossibility to test all possible subsets, even when the relevance criterion is simple to compute or estimate.

The distance measure, e.g., the Euclidean distance measure, is a very traditional discrimination or divergence measure. The dependence measure, also called the correlation measure, is mainly utilized to find the correlation between two features or a feature and a class. The consistency measure is heavily relied on the training data set and discussed for feature selection in Dash and Liu (2003). The measures are all sensitive to the concrete values of the training

data; hence they are easily affected by the noise or outlier data. Whereas the information measures, such as the entropy or mutual information, investigate the amount of information or the uncertainty of a feature for classification. The data classification process is aimed at reducing the amount of uncertainty or gaining information about the classification.

To perform feature subset validation, the available data may be further divided into training and validation sets. The search procedure is performed on the training set and the search stops when the stopping criterion, such as the significance test of the validation errors on the validation set, is satisfied. With the aim to ease the situation, Pudil et al. (2002) developed a software package for solving the Subset Selection problem.

The study of Hua et al. (2009) compares some basic feature-selection methods in settings involving thousands of features, using both model-based synthetic data and real data. Under this framework, it evaluates the performances of feature-selection algorithms for different distribution models and classifiers. Although the results clearly show that none of the considered feature-selection methods performs best across all scenarios, there are some general trends relative to sample size and relations among the features. For instance, the classifier-independent univariate filter methods have similar trends. Filter methods have better or similar performance with wrapper methods for harder problems. Wrapper methods have better performance when the sample size is sufficiently large. Moreover, Arauzo-Azofra et al. (2011) carry out an empirical evaluation of some feature selection methods applied in classification. A contraposition between accuracy and feature reduction is detected. This shows that methods performing greater reductions start losing relevant features, which leads them to worse accuracy results. For this reason, a single method cannot be recommended for all situations.

Some feature selection methods treat the multi-class case directly rather than decomposing it into several two-class problems. It is often argued that forward selection is computationally more efficient than backward elimination to generate nested subsets of features. However, the defenders of backward elimination argue that weaker subsets are found by forward selection because the importance of features is not assessed in the context of other features not included yet.

Therefore, in real applications, although feature ranking based on filters together with sequential forward search may yield suboptimal feature subsets, one may still choose hybrid techniques so as to take advantages of the filter and wrapper approaches.

References

Abe N, Kudo M (2006) Non-parametric classifier-independent feature selection. Pattern Recognit 39:737–746

Al-Ani A (2009) A dependency-based search strategy for feature selection. Expert Syst Appl 36:12392–12398

Ani A Al (2005) Feature subset selection using ant colony optimization. Int J Comput Intell 2(1):53–58

Arauzo-Azofra A, Aznarte JL, Benítez JM (2011) Empirical study of feature selection methods based on individual feature evaluation for classification problems. Expert Syst Appl. doi:10.1016/j.eswa.2010.12.160

Bacauskiene M, Verikasa A, Gelzinis A, Valincius D (2009) A feature selection technique for generation of classification committees and its application to categorization of laryngeal images. Pattern Recognit 42:645–654

Bae C, Yeh W-C, Chung YY, Liu S-L (2010) Feature selection with intelligent dynamic swarm and rough set. Expert Syst Appl 37:7026–7032

Bermejo P, de la Ossa L, Gámez JA, Puerta JM (2011) Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. Knowl Based Syst. doi:10.1016/j.knosys.2011.01.015

Bhavani SD, Rani TS, Bapi RS (2008) Feature selection using correlation fractal dimension: issues and applications in binary classification problems. Appl Soft Comput 8:555–563

Bonev B, Escalano F, Cazorla M (2008) Feature selection, mutual information, and the classification of high-dimensional patterns. Pattern Anal Appl 11(3–4):309–319
Caruana R, De Sa V (2003) Benefitting from the variables that variable selection discards. J Mach Learn Res 3:1245–1264
Chen W-C, Tseng S-S, Hong T-P (2008) An efficient bit-based feature selection method. Expert Syst Appl 34:2858–2869
Chen Y, Miao D, Wang R (2010) A rough set approach to feature selection based on ant colony optimization. Pattern Recognit Lett 31:226–233
Chow TWS, Huang D (2005) Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. IEEE Trans Neural Netw 16(1):213–224
Dash M, Liu H (2003) Consistency-based search in feature selection. Artif Intell 151(1–2):155–176
de Angelis V, Felici G, Mancinelli G (2006) Feature selection for data mining. In: Triantaphyllou E, F.elici G (eds) Data mining and knowledge discovery approaches based on rule induction techniques, massive computing series. Springer, Heidelberg pp 227–252
Drugan MM, Wiering MA (2010) Feature selection for Bayesian network classifiers using the MDL-FS score. Int J Approx Reason 51:695–717
ElAlami ME (2009) A filter model for feature subset selection based on genetic algorithm. Knowl Based Syst 22:356–362
Estévez PA (2009) Normalized mutual information feature selection. IEEE Trans Neural Netw 20(2):189–201
Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
Francois D, Rossi F, Wertz V, Verleysen M (2007) Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing 70:1276–1288
Fu X, Wang L (2005) Data mining with computational intelligence. Springer, Berlin
Gheyas IA, Smith LS (2010) Feature subset selection in large dimensionality domains. Pattern Recognit 43: 5–13
Gomez-Verdejo V, Verleysen M, Fleury J (2009) Information-theoretic feature selection for functional data classification. Neurocomputing 72:3580–3589
Guo B, Damper RI, Gunn SR, Nelson JDB (2008) A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. Pattern Recognit 41:1653–1662
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
Hall MA, Holmes G (2003) Benchmarking attribute selection techniques for discrete class data set mining. IEEE Trans Knowl Data Set Eng 15(3)
Hild II KE, Erdogmus D, Torkkola K, Principe JC (2006) Feature extraction using information theoretic learning. IEEE Trans Pattern Anal Mach Intell 28(9):1385–1392
Hong J-H, Cho S-B (2006) Efficient huge-scale feature selection with speciated genetic algorithm. Pattern Recognit Lett 27:143–150
Hsu H-H, Hsieh C-W, Lu M-D (2011) Hybrid feature selection by combining filters and wrappers. Expert Syst Appl. doi:10.1016/j.eswa.2010.12.156
Hu QH, Yu DR, Xie ZX, Liu JF (2006a) Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans Fuzzy Syst 14:191–201
Hu QH, Yu DR, Xie ZX (2006b) Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognit Lett 27:414–423
Hu Q, Liu J, Yu D (2008a) Mixed feature selection based on granulation and approximation. Knowl Based Syst 21:294–304
Hu Q, Yu D, Liu J, Wu C (2008b) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178:3577–3594
Hu Q, Che X, Zhang L, Yu D (2010) Feature evaluation and selection based on neighborhood soft margin. Neurocomputing 73:2114–2124
Hua J, Tembe WD, Dougherty ER (2009) Performance of feature-selectionmethods in the classification of high-dimension data. Pattern Recognit 42:409–424
Huang D, Chow TWS (2005) Effective feature selection scheme using mutual information. Neurocomputing 63:325–343
Huang J, Cai Y, Xu X (2007) A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognit Lett 28:1825–1844
Huang D, Gan Z, Chow TWS (2008) Enhanced feature selection models using gradient-based and point injection techniques. Neurocomputing 71:3114–3123
Jiao Na, Miao D, Zhou J (2010) Two novel feature selection methods based on decomposition and composition. Expert Syst Appl 37:7419–7426

Jin-Jie H, Ning L, Shuang-Quan L, Yun-Ze C (2008) Feature selection for classificatory analysis based on information-theoretic criteria. Acta Autom Sinica 34(3):383–392
Kabir Md M, Islam Md M, Murase K (2010) A new wrapper feature selection approach using neural network. Neurocomputing 73:3273–3283
Kahramanli S, Hacibeyoglu M, Arslan A (2011) A Boolean function approach to feature selection in consistent decision information systems. Expert Syst Appl. doi:10.1016/j.eswa.2011.01.002
Kalakech M, Biela P, Macaire L, Hamad D (2011) Constraint scores for semi-supervised feature selection: a comparative study. Pattern Recognit Lett 32:656–665
Kanan HR, Faez K (2008) An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system. Appl Math Comput 205:716–725
Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of the 1995 IEEE international conference on neural networkds, vol 4. Perth, Australia, pp 1942–1948
Kraskov A, Stogbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69
Kumar N, Kummamuru K (2008) Semisupervised clustering with metric learning using relative comparisons. IEEE Trans Knowl Data Eng 20:496–503
Lai C, Reinders MJT, Wessels L (2006) Random subspace method for multivariate feature selection. Pattern Recognit Lett 27:1067–1076
Lashkia G, Anthony L (2004) Relevant, irredundant feature selection and noisy example elimination. IEEE Trans Syst Man Cybern B Cybern 34(2):888–897
Lee C, Lee GG (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manag 42(1):155–165
Lee C-P, Leu Y (2011) A novel hybrid feature selection method for microarray data analysis. Appl Soft Comput 11:208–213
Lee S-K, Yi S-J, Zhang B-T (2006) Combining information-based supervised and unsupervised feature selection. StudFuzz, 489–498
Li Y, Lu B-L (2009) Feature selection based on loss-margin of nearest neighbor classification. Pattern Recognit 42:1914–1921
Li G-Z, Yang JY (2008) Feature selection for ensemble learning and its application. In: Zhang Y-Q, Rajapakse JC (eds) Machine learning in bioinformatics. Wiley, New York
Liang J, Yang Su, Winstanley A (2008) Invariant optimal feature selection: a distance discriminant and feature ranking based solution. Pattern Recognit 41:1429–1439
Liu Y, Zheng YF (2006) FS_SFS: anovel feature selection method for support vector machines. Pattern Recognit 39:1333–1345
Liu H, Motoda H (2008) Computational methods of feature selection. Chapman & Hall/CRC, London
Liu H, Sun J, Liu L, Zhang H (2009) Feature selection with dynamic mutual information. Pattern Recognit 42:1330–1339
Liu H, Liu L, Zhang H (2009b) Boosting feature selection using information metric for classification. Neurocomputing 73:295–303
Lutu PEN, Engelbrecht AP (2010) A decision rule-based method for feature selection in predictive data mining. Expert Syst Appl 37:602–609
Luukka P (2011) Feature selection using fuzzy entropy measures with similarity classifier. Expert Syst Appl 38:4600–4607
Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. Inf Sci 179:2208–2217
Malhi A, Gao RX (2004) PCA-based feature selection scheme for machine defect classification. IEEE Trans Instrum Meas, 1517–1525
Mao KZ (2004) Orthogonal forward selection and backward elimination algorithms for feature subset selection. IEEE Trans Syst Man Cybern B Cybern 34(1):629–634
Meyer PE, Schretter C, Bontempi G (2008) Information-theoretic feature selection in microarray data using variable complementarity. IEEE J Sel Top Signal Process 2(3):261–274
Moustakidis SP, Theocharis JB (2010) SVM-FuzCoC: a novel SVM-based feature selection method using a fuzzy complementary criterion. Pattern Recognit 43:3712–3729
Nakariyakul S, Casasent DP (2009) An improvement on floating search algorithms for feature subset selection. Pattern Recognit 42:1932–1940
Ng WWY, Yeung DS, Firth M, Tsang ECC, Wang X-Z (2008) Feature selection using localized generalization error for supervised classification problems using RBFNN. Pattern Recognit 41:3706–3719
Nguyen MH, de la Torre F (2010) Optimal feature selection for support vector machines. Pattern Recognit 43:584–591
Oh I-S, Lee J-S, Moon B-R (2004) Hybrid genetic algorithms for feature selection. IEEE Trans Pattern Anal Mach Intell 26(11):1424–1437

Parpinelli RS, Lopes HS, Freitas AA (2002) Data mining with an ant colony optimization algorithm. IEEE Trans Evol Comput 6:321–332
Pawlak Z (1991) Rough sets—theoretical aspects of reasoning about data. Kluwer, Dordrecht
Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency max-relevance and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
Peng Y, Wu Z, Jiang J (2010) A novel feature selection approach for biomedical data classification. J Biomed Inf 43:15–23
Pineda-Bautista BB, Carrasco-Ochoa JA, Martınez-Trinidad JF (2011) General framework for class-specific feature selection. Expert Syst Appl. doi:10.1016/j.eswa.2011.02.016
Piramuthu S (2004) Evaluating feature selection methods for learning in data mining applications. Eur J Oper Res 156:483–494
Plumbley MD, Oja E (2004) A nonnegative PCA algorithm for independent component analysis. IEEE Trans Neural Netw 15(1):66–76
Pudil P, Novoviov J, Somol P (2002) Feature selection toolbox software package. Pattern Recognit Lett 23:487–492
Qu G, Hariri S, Yousif M (2005) A new dependency and correlation analysis for features. IEEE Trans Knowl Data Eng 17(9):1199–1207
Reunanen J (2003) Overfitting in making comparisons between variable selection methods. J Mach Learn Res 3:1371–1382
Saethang T, Prom-on S, Meechai A, Chan JH (2009) Sample filtering relief algorithm: robust algorithm for feature selection. ICONIP 2008, Part II, LNCS 5507, pp 260–267
Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517
Salamó M, López-Sánchez M (2011) Rough set based approaches to feature selection for case-based reasoning classifiers. Pattern Recognit Lett 32:280–292
Sánchez L, Rosario Suárez M, Villar JR, Couso I (2008) Mutual information-based feature selection and partition design in fuzzy rule-based classifiers from vague data. Int J Approx Reason 49:607–622
Sindhwani V, Rakshit S, Deodhare D, Erdogmus D, Principe J, Niyogi P (2004) Feature selection in MLPs and SVMs based on maximum output information. IEEE Trans Neural Netw 15(4):937–948
Sivagaminathan RK, Ramakrishnan S (2007) A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert Syst Appl 33:49–60
Somol P, Pudil P, Kittler J (2004) Fast branch and bound algorithms for optimal feature selection. IEEE Trans Pattern Anal Mach Intell 26(7):900–912
Sotoca JM, Pla F (2010) Supervised feature selection by clustering using conditional mutual information-based distances. Pattern Recognit 43:2068–2081
Stoppiglia H, Dreyfus G, Dubois R, Oussar Y (2003) Ranking a random feature for variable and feature selection. J Mach Learn Res 3:1399–1414
Sun Y (2007) Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell 29:1035–1051
Sun D, Zhang D (2010) Bagging constraint score for feature selection with pairwise constraints. Pattern Recognit 43:2106–2118
Swiniarski RW, Skowron A (2003) Rough set methods in feature selection and recognition. Pattern Recognit Lett 24:833–849
Tahir MA, Bouridane A, Kurugollu F (2007) Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recognit Lett 28:438–446
Tang W, Mao KZ (2007) Feature selection algorithm for mixed data with both nominal and continuous features. Pattern Recognit Lett 28:563–571
Uncu O, Turksen IB (2007) A novel feature selection approach: combining feature wrappers and filters. Inf Sci 177(2):449–466
Unler A, Murat A, Chinnam RB (2010) mr2PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Inf Sci. doi:10.1016/j.ins.2010.05.037
Verikas A, Bacauskiene M, Valincius D, Gelzinis A (2008) Predictor output sensitivity and feature similarity-based feature selection. Fuzzy Sets Syst 159:422–434
Wang X, Yang J, Teng X, Xia W, Jensen R (2007) Feature selection based on rough sets and particle swarm optimization. Pattern Recognit Lett 28:459–471
Wang L, Zhou N, Chu F (2008) A general wrapper approach to selection of class- dependent features. IEEE Trans Neural Netw 19(7):1267–1278
Wang C-M, Huang Y-F (2009) Evolutionary-based feature selection approaches with new criteria for data mining: a case study of credit approval data. Expert Syst Appl 36:5900–5908

Wang Y, Li L, Ni J, Huang S (2009) Feature selection using tabu search with long-term memories and probabilistic neural networks. Pattern Recognit Lett 30:661–670

Xie J, Wang C (2011) Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Syst Appl 38:5809–5815

Yang J, Olafsson S (2006) Optimization-based feature selection with adaptive instance sampling. Comput Oper Res 33:3088–3106

Yang M, Song J (2010) A novel hypothesis-margin based approach for feature selection with side pairwise constraints. Neurocomputing 73:2859–2872

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

Yusta SC (2009) Different metaheuristic strategies to solve the feature selection problem. Pattern Recognit Lett 30:525–534

Zhang D, Chen S, Zhou Z-H (2008) Constraint score: a new filter method for feature selection with pairwise constraints. Pattern Recognit 41:1440–1451

Zhao J, Lu K, He X (2008) Locality sensitive semi-supervised feature selection. Neurocomputing 71: 1842–1849

Zheng H, Zhang Y (2008) Feature selection for high-dimensional data in astronomy. Adv Space Res 41: 1960–1964

Zhu ZX, Ong Y-S, Dash M (2007) Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans Syst Man Cybern B Cybern 37(1):70–76