UNIVERSITY INSTITUTE OF TECHNOLOGY
THE UNIVERSITY OF BURDWAN

Seminar Report - II

IT-892(OLD)

Feature Selection Techniques

Prakash Anand
4th year, B.E. (I.T.)
Roll No.: 20163016
Reg. Number: A4516 of 2016-17

Acknowledgement

I thank GOD, the almighty for giving me strength and knowledge to do this .I would like to thank and deep sense of gratitude to our guide Ms. Kasturi Ghosh, Professor and In-charge of the Department of Information Technology, for his valuable advice and suggestions as well as his continued guidance, patience and support that helped us to shape and refine our work.I would like to extend our sincere thanks to all the teaching and non-teaching staff of our department who have contributed directly and indirectly during the course .

Finally, I would like to thank our parents and friends for their patience, cooperation and moral support throughout our life.

Prakash Anand

Abstract

Feature Selection is an essential component of machine learning and data mining which has been studied for many years under many different conditions and in diverse scenarios. These algorithms aim at ranking and selecting a subset of relevant features according to their degrees of relevance, preference, or importance as defined in a specific application. Because feature selection can reduce the amount of features used for training classification models, it alleviates the effect of the curse of dimensionality, speeds up the learning process, improves model's performance, and enhances data understanding.Using FS methods not only reduces the burden of the data but also avoids overfitting of the model.

Table of Contents

Introduction

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Adequate selection of features may improve accuracy and efficiency of classifier methods. Feature ranking and selection for classification aims at reducing the dimensionality and noise in data sets. FS performs information filtering since it removes redundant or unwanted information from an information stream. Feature selection improves algorithms performance and classification accuracy since the chance of overfitting increases with the number of features.

Objectives and Scope

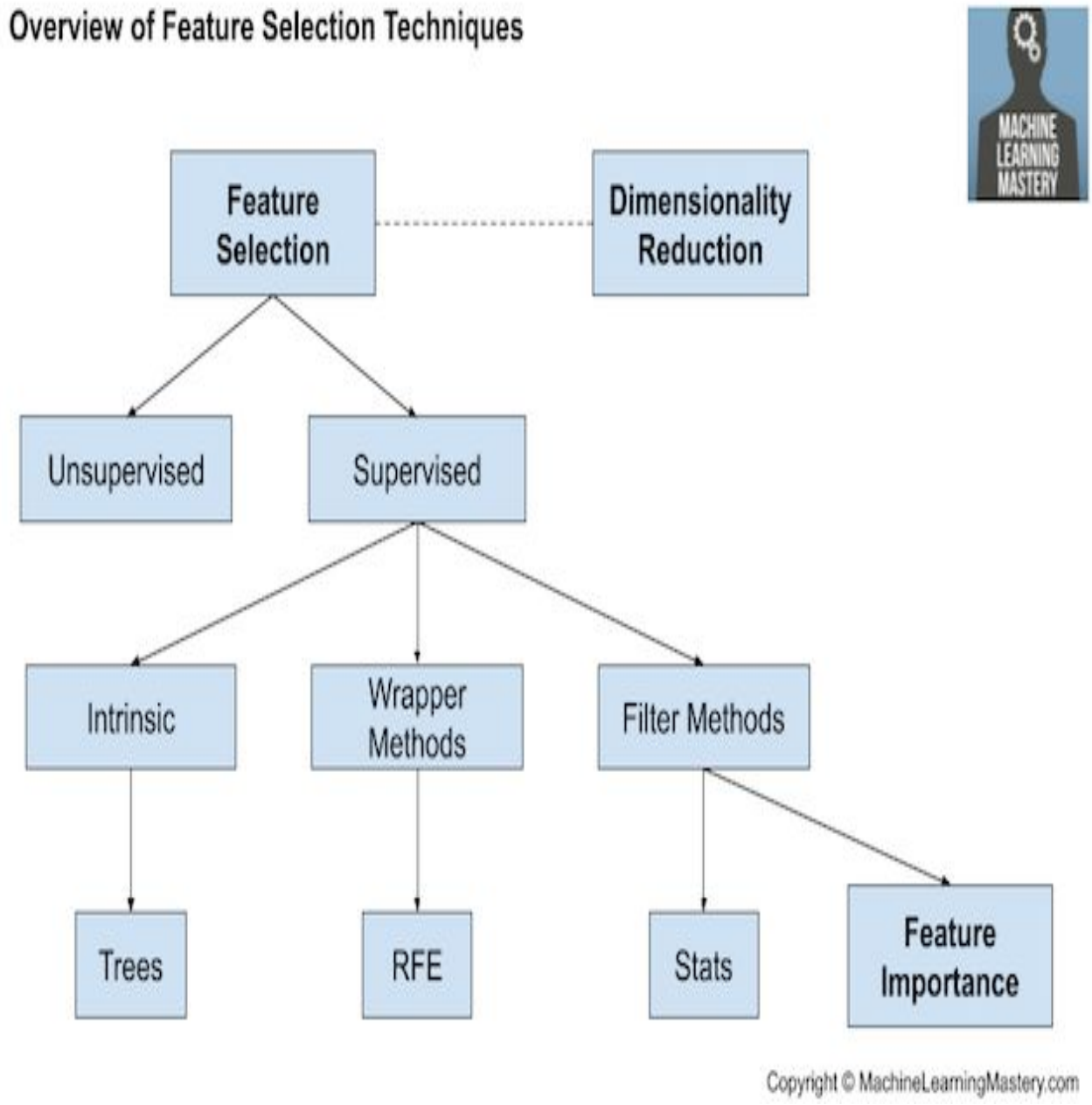
The objective of feature selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.Irrelevant features can be removed without affecting learning performance.feature selection is a widely recognized important task in machine learning, artificial intelligence, computer vision, and data mining successfully applied in fields like information retrieval (e.g., feature-based user retrieval), user re-identification by soft-biometrics, recommendation systems, visual object tracking for real-time feature ranking and selection and many other domains.

Need for feature selection

Benefits of performing feature selection before modeling your data:

- **Reduces Overfitting**:Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy**: Less misleading data means modeling accuracy improves.
- **Reduces Training Time**: fewer data points reduce algorithm complexity and algorithms train faster.
- To improve the classifier by removing the irrelevant features and noise.
- To identify the relevant features for any specific problem.
- To improve the performance of learning algorithms.
- Reduction in features to improve the quality of prediction.
- Reduces the size of the problem.

Overview of Feature Selection

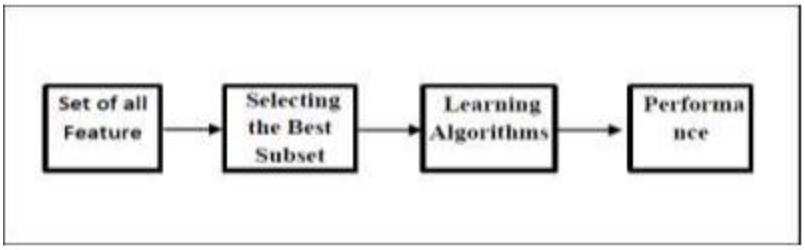


Feature Selection Algorithms

Different kinds of feature selection algorithms have been proposed.The feature selection techniques are categorized into five methods namely Filter methods, Wrapper method, Embedded method,online-based and hybrid-based. Every feature selection algorithm uses any one of the three feature selection techniques.The two search algorithms forward selection and backward eliminations are used to select and eliminate the appropriate feature. The feature selection is a three step process namely search, evaluate and stop.In feature extraction, some attributes of the existing data, intended to be informative, are extracted.

Filter Method

This method selects the feature without depending upon the type of classifier used. The advantage of this method is that it is simple and independent of the type of classifier used so feature selection needs to be done only once and the drawback of this method is that it ignores the interaction with the classifier, ignores the feature dependencies, and lastly each feature considered separately.This method has acceptable stability and scalability, and can also be used in offline feature selection applications.



Filter Approach

Univariate Filter Method

Univariate Feature Selection is a statistical method used to select the features which have the strongest relationship with our correspondent labels. Using the SelectKBest method we can decide which metrics to use to evaluate our features and the number of K best features we want to keep.Univariate feature selection examines each feature individually to determine the strength of the relationship of the feature with the response variable. These methods are simple to run and understand and are in general particularly good for gaining a better understanding of data.There are a lot of different options for univariate selection such as Pearson Correlation,Distance Correlation,Model based ranking etc.

Multivariate Filter Method

Multivariate feature selection methods can avoid this problem by computing multivariate statistics for feature ranking because they consider the dependencies between the features when calculating scores for features. The most common search strategies that can be used with multivariate filters can be categorized into exponential algorithms, sequential algorithms and randomized algorithms.As an information-based multivariate feature selection method that aims to pick features containing information about the experimental condition, searchlight is sensitive to features that might be discarded by univariate methods.

Types of Filter Method

The filter attribute selection method is independent of the classification algorithm. Filter method is further categorized into two types:

1. Attribute evaluation algorithms
2. Subset evaluation algorithms

The algorithms are categorized based on whether they rate the relevance of individual features or feature subsets.Attribute evaluation algorithms rank the features individually and assign a weight to each feature according to each feature's degree of relevance to the target feature. The attribute evaluation methods are likely to yield subsets with redundant features since these methods do not measure the correlation between features. Subset evaluation methods, in contrast, select feature subsets and rank them based on certain evaluation criteria and hence are more efficient in removing redundant features.

Pearson Correlation

One of the simplest methods for understanding a feature's relation to the response variable is Pearson Correlation Coefficient, which measures linear correlation between two variables. The resulting value lies in [-1;1], with -1 meaning perfect negative correlation (as one variable increases, the other decreases), +1 meaning perfect positive correlation and 0 meaning no linear correlation between the two variables.It's fast and easy to calculate and is often the first thing to be run on the data. Scipy's pearsonr method computes both the correlation and p-value for the correlation, roughly showing the probability of an uncorrelated system creating a correlation value of this magnitude.

Chi-Square Test

It is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.The chi-squared test is one of the feature selection methods used in the filter method. The chi squared statistical test checks the independence between the two events. If X, Y are two events then the statistical independence is denoted by the following equations:

$$P(XY) = P(X)\,P(Y) \text{ or}$$

$$P(X/Y) = P(X) \text{ and } P(Y/X) = P(Y)$$

The main disadvantage of the filter method is it ignores the dependencies among the features and treats the features individually.

Distance correlation

Another robust, competing method of correlation estimation is distance correlation, designed explicitly to address the shortcomings of Pearson correlation. While for Pearson correlation, the correlation value 0 does not imply independence distance correlation of 0 does imply that there is no dependence between the two variables.Distance correlation provides a new approach to the problem of testing the joint independence of random vectors. For all distributions with finite first moments, distance correlation R generalizes the idea of correlation in two fundamental ways:

(i) R(X, Y) is defined for X and Y in arbitrary dimensions;
(ii) R(X, Y) = 0 characterizes independence of X and Y.

This can relay useful extra information on whether the relationship is negative or positive, i.e. do higher feature values imply higher values of the response variables or vice versa.

Fast Correlation based Feature Selection

FCBF (Fast Correlation Based Filter) [4] is a multivariate feature selection method which starts with a full set of features, uses symmetrical uncertainty to calculate dependencies of features and finds the finest subset using backward selection technique with sequential search strategy. The FCBF algorithm consists of two stages: the first one is a relevance analysis that orders the input variables depending on a relevance score, which is computed as the symmetric uncertainty with respect to the target output. This stage is also used to discard irrelevant variables, whose ranking score is below a predefined threshold. The second stage is a redundancy analysis, which selects predominant features from the relevant set obtained in the first stage. This selection is an iterative process that removes those variables which form an approximate Markov blanket.

Information Gain

Information gain tells us how important a given attribute of the feature vectors is. IG feature selection method selects the terms having the highest information gain scores. Information gain measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature) defined as:

$$Entropy = -pi\ log2pi\ n\ i=1$$

where _n_ is the number of classes, and the Pi is the probability of S belongs to class _i_. The gain of A and S is calculated as:

$$Gain\ A = Entropy\ S - \mathbf{Sk\ S} * \mathbf{Entropy\ Sk}\ m\ k=1 \quad (5)$$

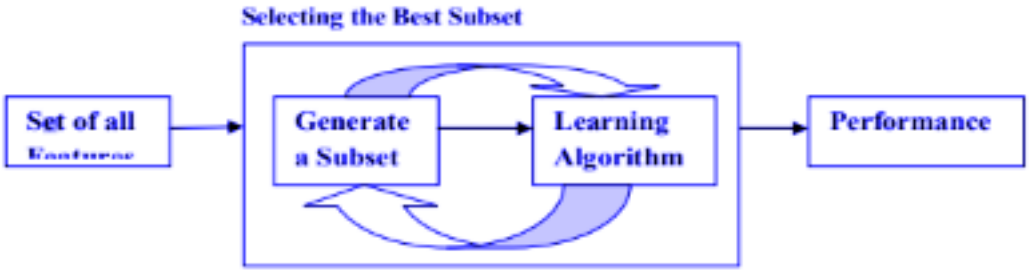
Where, Sk is the subset of S.

Common filter method

Name	Filter class	Applicable to task
Information gain	univariate, information	classification
Gain ratio	univariate, information	classification
Symmetrical uncertainty	univariate, information	classification
Correlation	univariate, statistical	regression
Chi-square	univariate, statistical	classification
Inconsistency criterion	multivariate, consistency	classification
Minimum redundancy, maximum relevance (mRmR)	multivariate, information	Classification regression
Correlation-based feature selection (CFS)	multivariate, statistical	Classification regression
Fast correlation-based filter (FCBF)	multivariate, information	classification
Fisher score	univariate, statistical	classification
Relief and ReliefF	univariate, distance	classification
Spectral feature selection (SPEC) and Laplacian Score (LS)	univariate, similarity	classification
Feature selection for sparse clustering	multivariate, similarity	clustering
Localized Feature Selection Based on Scatter Separability (LFSBSS)	multivariate, statistical	clustering
Multi-Cluster Feature Selection (MCFS)	multivariate, similarity	clustering
Feature weighting Kmeans	multivariate, statistical	clustering
ReliefC	univariate, distance	clustering

Wrapper methods

In this method the feature is dependent upon the classifier used, i.e. it uses the result of the classifier to determine the goodness of the given feature or attribute. The advantage of this method is that it removes the drawback of the filter method, i.e. It includes the interaction with the classifier and also takes the feature dependencies and drawback of this method is that it is slower than the filter method because it takes the dependencies also. The quality of the feature is directly measured by the performance of the classifier.



Wrapper Approach

Types of wrapper methods

Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

- **Forward Selection**

  Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

- **Backward Elimination**

  In backward elimination, we start with all the features and remove the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

- **Recursive Feature elimination:** It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

- **Plus L minus R (LRS)**

  The Plus L is a generalization of SFS and the Minus R is the generalization of SBE algorithms. If L>R then The algorithm starts with the empty set and adds the necessary features to the resultant set, else the algorithm starts with the entire set and starts eliminating the irrelevant features and produces the resultant set.

- **Simulated annealing**

  The simulated annealing is a method which is used for optimization problems. Initially the solution S and the related cost function F(c) is given as input values, and the algorithm is executed to find out the resultant solution S' and the minimum cost function F'(c). The algorithm proceeds till the optimum result.

- **Randomized hill climbing**

  The hill climbing is a local search algorithm.

  Step1.procedure hill-climbing (Max_Flips)
  Step2.restart: s <- random valuation variables;
  Step 3.for j: =1 to Max_Flips do
  Step4.if eval(s) =0 then return s endif;
  Step5.if s is a strict local minimum then
  Step 6.goto restart
  Step7.else
  Step 8.s <- neighborhood with smallest evaluation value
  Step 9.endif
  Step 10.endfor
  Step 11.goto restart
  Step12.end

Wrapper Method Algorithm

Sequential Selection Algorithms

The Sequential Feature Selection (SFS) [7][8][9] algorithm starts with an empty set and adds one feature for the first step which gives the highest value for the objective function. After the first step, the remaining features are added individually to the current subset and the new subset is evaluated. The individual features that give maximum classification accuracy are permanently included in the subset. The process is repeated until we get the required number of features. This algorithm is called a naive SFS algorithm since the dependency between the features is not taken into consideration.

A Sequential Backward Selection algorithm is exactly the reverse of the SFS algorithm. Initially, the algorithm starts from the entire set of variables and removes one irrelevant feature at a time whose removal gives the lowest decrease in predictor performance.

Wrapper Method Algorithm

Heuristic Search Algorithms

Heuristic search algorithms include Genetic algorithms (GA)[12], Ant Colony Optimization(ACO)[13], Particle Swarm Optimization(PSO)[14],etc. A genetic algorithm is a search technique used in computing to find true or approximate solutions to optimization and search problems. Genetic algorithms are based on the Darwinian principle of survival of the fittest theory. ACO is based on the shortest paths found by real ants in their search for food sources. ACO approaches suffer from inadequate rules of pheromone update and heuristic information.They do not consider random phenomenon of ants during subset formations. PSO approach does not employ crossover and mutation operators, hence is efficient over GA but requires several mathematical operators.

Filter vs Wrapper methods

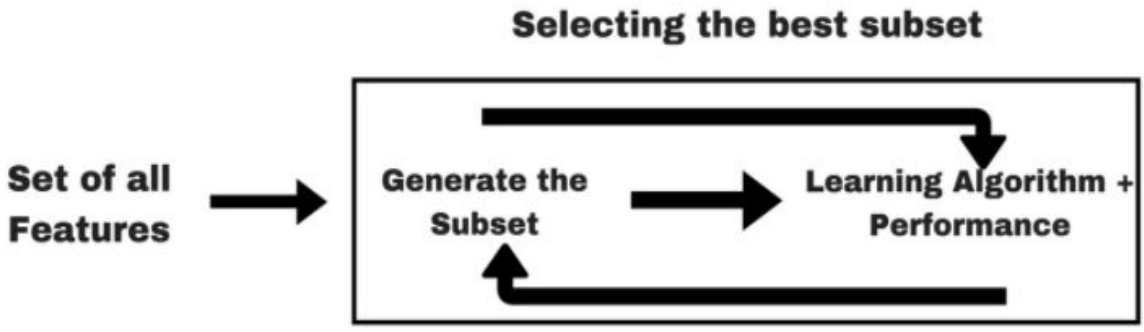
The main differences between the filter and wrapper methods for feature selection are:

- Filter methods measure the relevance of features by their correlation with dependent variables while wrapper methods measure the usefulness of a subset of features by actually training a model on it.
- Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
- Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
- Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features.
- Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.

Embedded Methods

Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting.

- Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.
- Ridge regression performs L2 regularization which adds a penalty equivalent to square of the magnitude of coefficients.



Embedded Methods

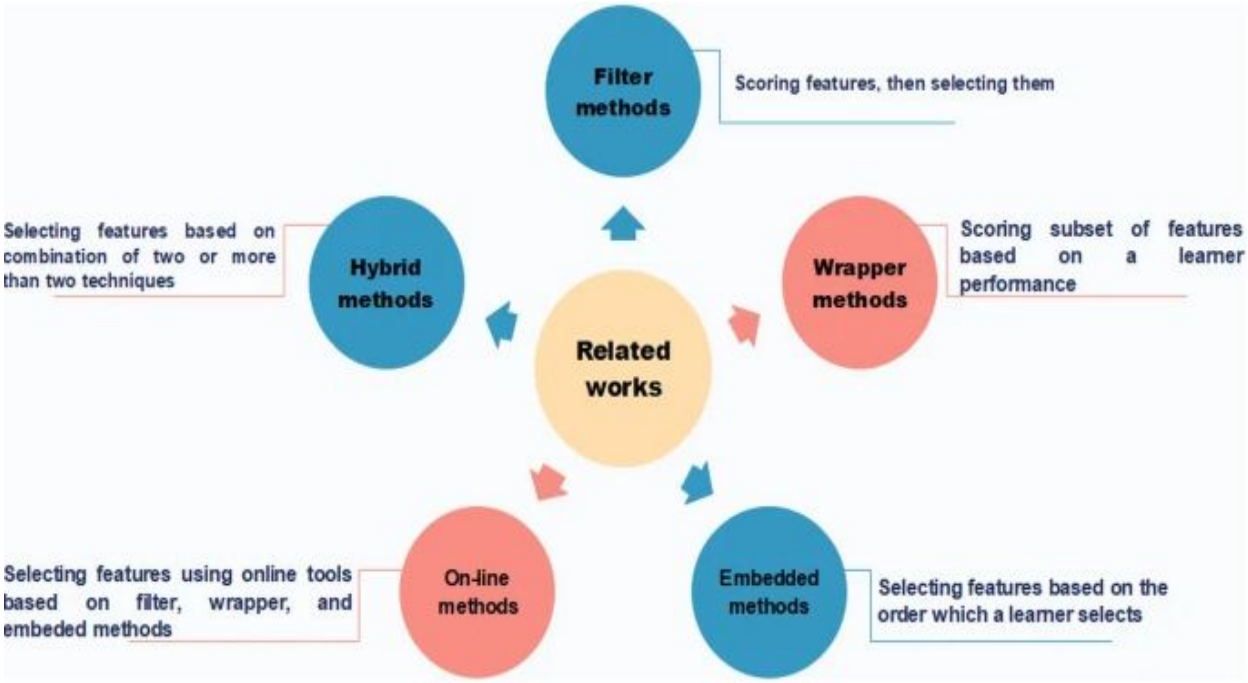
Online-based Method

These methods select features using online user tips. In a related work, a feature cluster taxonomy feature selection (FCTFS) method has been introduced. The main goal of FCFS is the selection of features based on a user-guided mode. The accuracy of this method is lower than that of the other methods. In a separate study, an online feature selection method based on the dependency on the k nearest neighbours (k-OFSD) has been proposed, and this is suitable for high-dimensional datasets. The main motivation for the abovementioned work is the selection of features with a higher ability to separate those for which the performance has been examined using unbalanced data.A library of online feature selection (LOFS) has also been developed using the state-of-art algorithms, for use with MATLAB and OCTAVE. Since the performance of LOFS has not been examined for a range of datasets, its performance has not been investigated.

Hybrid-based Method

These methods are a combination of four above categories. For example, some related works use two-step feature selection methods. In these methods, a number of features are reduced by the first method, and the second method is then used for further reduction. While some works focus on only one of these categories, a hybrid two-step feature selection method, which combines the filter and wrapper methods, has been proposed for multi-word recognition. It is possible to remove the most discriminative features in the filter method, so that this method is solely dependent on the filter stage. DNA microarray datasets usually have a large size and a large number of features, and feature selection can reduce the size of this dataset, allowing a classifier to properly classify the data. For this purpose, a new hybrid algorithm has been suggested that combines the maximisation of mutual information with a genetic algorithm.

Application of feature Selection



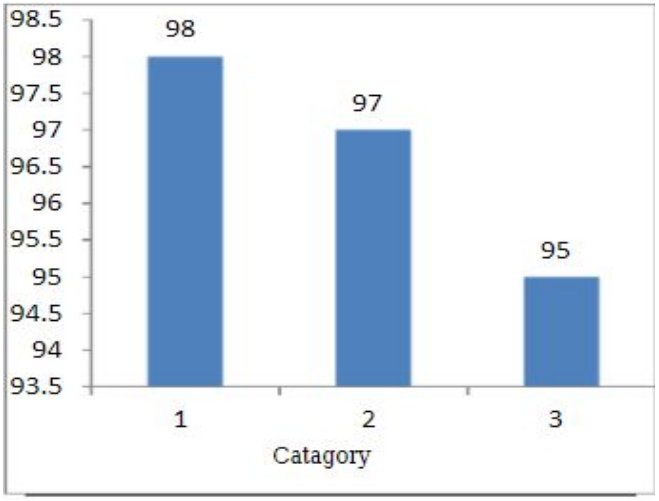
Classification of the related works. They have been categorized into five classes, including: (i) Filter method which scores features and then selects them. (ii) Wrapper method which scores a subset of features based on a learner performance. (iii) Embedded method which selects features based on the order that a learner selects them. (iv) Online method which is based on online tools. (V) Hybrid method which combines different methods in order to acquire better results.

Analysis of feature selection

The existing study shows that most of the feature selection algorithms are applied on synthetic dataset and few of them use the real data.The analysis result concludes that when the features are increased the accuracy of an algorithm gets affected and is decreased. It is a challenging task to keep the accuracy at the highest level even if the features are increased.

Average accuracy of feature selection algorithm and the maximum number of attributes

Category	No of Attributes	Accuracy In %
1	10-99	98
2	100-999	97
3	1000-9999	95



Mean accuracy of the feature selection algorithm with maximum number of features

Conclusion

Feature selection is an important issue in classification, because it may have a considerable effect on accuracy of the classifier. It reduces the number of dimensions of the dataset, so the processor and memory usage reduce; the data becomes more comprehensible and easier to study on. In this study, various feature selection techniques have been discussed and among the three approaches to feature selection method, filter methods should be used to get results in lesser time and for large datasets. If the results are accurate and optimal, a wrapper method like GA should be used.This study of feature selection algorithms of large surveys shows that the feature selection algorithm consistently improves the accuracy of the classifier.The feature selection algorithm must select the relevant features and also remove the irrelevant and inconsistent features which cause the degradation of accuracy of the classification algorithms.

Future scope

The future improvement from this study is to design a feature selection algorithm for high dimensional multiclass dataset with considerable improvement in accuracy with less space and time requirement.From a data mining perspective, a multiple disturbance classification problem can be treated as a multi-label classification problem where a single event is tagged with the label of multiple disturbances. Also,I will use factor analysis techniques to reduce a large number of variables into fewer number of factors. By using this technique we extract maximum common variance from all variables and put them into a common score. As an index of all variables, we will use this score for further analysis. There are also a wide range of other multi-label algorithms that can be experimented with to classify multiple disturbances.

References

[1] Asha Gowda Karegowda, M.A.Jayaram and A.S. Manjunath, —Feature Subset Selection Problem using Wrapper Approach in Supervised Learning || , International Journal of Computer Applications, Vol. 1, No. 7, pp. 0975–8887, 2010.

[2] Ron Kohavi, George H. John, —Wrappers for feature subset Selection || , Artificial Intelligence, Vol. 97, No. 1-2. pp. 273-324, 1997.

[3] S. Doraisami, S. Golzari, A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, Content-Based Retrieval, Categorization and Similarity, 2008

[4] A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benítez, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications, 38 (2011) 8170-8177.

[5] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International Journal of Engineering Research & Technology (IJERT).