A Review on Feature Selection Methods for DataMining

Mitul S Patel
Department of computer science & engineering
Parul institute of Technology,Vadodara
mitul.patel.8692@gmail.com

Asst. Prof. Dinesh Vaghela
Department of computer science & engineering
Parul institute of Technology,waghodiya,Vadodara

ABSTRACT

Feature selection (FS) methods can be used in data pre-processing to achieve efficient data reduction. This is useful for finding accurate data models. Since exhaustive search for optimal feature subset is infeasible in most cases, many search strategies have been proposed in literature. The usual applications of FS are in classification, clustering, and regression tasks. This review considers most of the commonly used FS techniques. Particular emphasis is on the application aspects. In addition to standard filter, wrapper, and embedded methods, we also provide insight into FS for recent hybrid approaches and other advanced topics.

KEY WORDS : *Feature Selection; Filter Method; Wrapped Method; Embedded Method; Hybrid Method.*

1.INTRODUCTION

Feature selection is used to satisfy the common goal of maximizing the accuracy of the classifier, minimizing the related measurement costs; improve accuracy by reducing irrelevant and possibly redundant features; reduce the complexity and the associated computational cost; and improve the probability that a solution will be comprehensible and realistic[13]. Feature selection is one of the stage for preprocessing the data to reduce the dimensionality[10]. It selects a subset of the existing features without any transformation. It can be said as a special case of feature extraction.

In this paper, we focus on feature selection and provide an overview of the existing methods that are available for handling several different classes of problems. Additionally, we consider the most important application domains and review comparative studies on feature selection therein, in order to investigate which methods perform best for specific tasks. This research is motivated by the fact that there is an abundance of work in this field and insufficient systematization, particularly with respect to various application domains and novel research topics.

When we use feature selection, smaller number of features are extracted which means fewer model parameters. It improves the generalization capabilities and reduces complexity and execution time. Feature Selection methodologies can be categorized as: supervised and unsupervised feature selection methods.

2.FEATURE SELECTION METHODS

Feature selection methods can be classified in a number of ways. The most common one is the classification into filters, wrappers, embedded, and hybrid methods[5]. The abovementioned classification assumes feature independency or near-independency. Additional methods have been devised for datasets with structured features where dependencies exist and for streaming features[4].

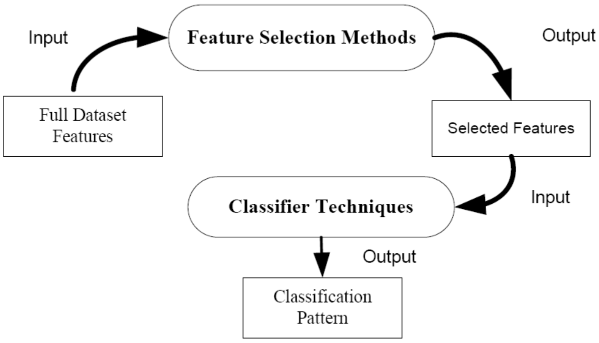
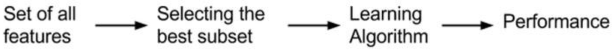


FIGURE 1 : Feature Selection for Classification

A. Filter Method :



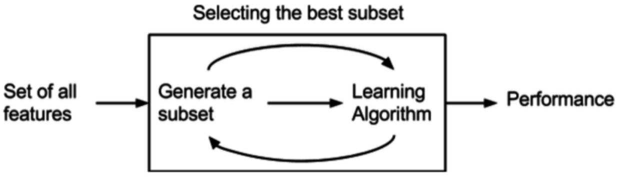
Filter methods select features based on a performance measure regardless of the employed data modeling algorithm. Only after the best features are found, the modeling algorithms can use them. Filter methods can rank individual features or evaluate entire feature subsets. We can roughly classify the developed measures for feature filtering into: information, distance, consistency, similarity, and statistical measures. While there are many filter methods described in literature, a list of common methods is given in Table I, along with the appropriate references that provide details. Not all the filter features can be used for all classes of data mining tasks. therefore, the filters are also classified depending on the task: classification, regression or clustering. Due to lack of space, we do not consider semi-supervised learning feature selection methods in this work.

Univariate feature filters evaluate (and usually rank) a single feature, while multivariate filters evaluate an entire feature subset. Feature subset generation for multivariate filters depends on the search strategy. While there are many search strategies, there are four usual starting points for feature subset generation:

1. forward selection
2. backward elimination
3. bidirectional selection
4. Heuristic feature subset selection.

Forward selection typically starts with an empty feature set and then considers adding one or more features to the set. Backward elimination typically starts with the whole feature set and considers removing one or more features from the set. Bidirectional search starts from both sides - from an empty set and from the whole set, simultaneously considering larger and smaller feature subsets. Heuristic selection generates a starting subset based on a heuristic (e.g. a genetic algorithm), and then explores it further.

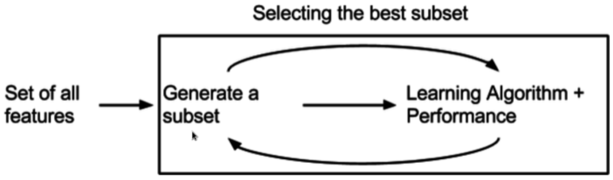
B. Wrapped Method



Wrappers consider feature subsets by the quality of the performance on a modeling algorithm, which is taken as a black box evaluator. Thus, for classification tasks, a wrapper will evaluate subsets based on the classifier performance (e.g. Naïve Bayes or SVM) , while for clustering, a wrapper will evaluate subsets based on the performance of a clustering algorithm (e.g. *K*-means). The evaluation is repeated for each subset, and the subset generation is dependent on the search strategy, in the same way as with filters. Wrappers are much slower than filters in finding sufficiently good subsets because they depend on the resource demands of the modeling algorithm. The feature subsets are also biased towards the modelling algorithm on which they were evaluated (even when using cross-validation). Therefore, for a reliable generalization error estimate, it is necessary that both an independent validation sample and another modeling algorithm are used after the final subset is found. On the other hand, it has been empirically proven that wrappers obtain subsets with better perfomance than filters because the subsets are evaluated using a real modeling algorithm.

Practically any combination of search strategy and modelling algorithm can be used as a wrapper, but wrappers are only feasible for greedy search strategies and fast modelling algorithms such as Naïve Bayes, linear SVM, and Extreme Learning Machines.

C.Embedded Method



Embedded methods perform feature selection during the modeling algorithm's execution. These methods are thus embedded in the algorithm either as its normal or extended functionality. Common embedded methods include various types of decision tree algorithms: CART, C4.5, random forest, but also other algorithms (e.g. multinomial logistic regression and its variants). Some embedded methods perform feature weighting based on regularization models with objective functions that minimize fitting errors and in the mean time force the feature coefficients to be small or to be exact zero. These methods based on Lasso or Elastic Net usually work with linear classifiers (SVM or others) and induce penalties to features that do not contribute to the model.

D.Hybride Method

Hybrid methods were proposed to combine the best properties of filters and wrappers. First, a filter method is used in order to reduce the feature space dimension space, possibly obtaining several candidate subsets. Then, a wrapper is employed to find the best candidate subset. Hybrid methods usually achieve high accuracy that is characteristic to wrappers and high efficiency characteristic to filters. While practically any combination of filter and wrapper can be used for constructing the hybrid methodology, several interesting methodologies were recently proposed, such as: fuzzy random forest based feature selection, hybrid genetic algorithms, hybrid ant colony optimization, or mixed gravitational search algorithm.

3. APPLICATION OF FEATURE SELECTION IN VARIOUS FIELD

Text classification:
There are various challenges related to automated text classification such as: 1. An appropriate data structure is to be selected to represent the documents. 2. An appropriate objective function is to be chosen to optimize to avoid overfitting and obtain good generalization. Along with it algorithmic issues arising as a result of the high formal dimensionality of the data are to be dealt with [1].

Genre classification:
Metadata such as filename, author, size, date, track length and genres are the common features used to classify and retrieve genre documents. On the basis of these data, the classification is infeasible, so the feature selection step is required. In case of genre classification feature selection is a process where a segment of an audio is characterized into a compact numerical representation [2]. Feature selection is done to reduce the dimensionality of the data as a preprocessing step prior to classification due to high dimensionality of the feature sets.

Microarray data analysis:
1. Almost all bioinformatics problems have the number of features significantly larger than the number of samples (high feature to sample ratio datasets) [3]. For example: Breast cancer classification on the basis of microarray data. Though the information about all the genes is not required in case of such classification.

2. Content analysis and signal analysis in genomics also require feature selection.

Software defect prediction:
There are various software quality assurance attributes such as reliability, functionality, fault proneness, reusability, comprehensibility etc [6]. It is a critical issue to select most appropriate software metrics that likely to indicate fault proneness.

Sentiment analysis:
Sentiment analysis is capturing favorability using natural language processing. It is not just a topic based categorization. It deals with the computational treatment of opinion, sentiment, and subjectivity in text. It is useful in recommendation systems and question answering [8]. To decide about the positivity or negativity of the opinion on the basis of various features such as term presence, feature frequency, feature presence, term position, POS tags, syntax, topic and negation etc. All of the features are not required in each and every case. So feature selection need to be performed.

Stock market analysis:
There are hundreds of stock index futures. Along with it financial data including the stock market data is too complex to be searched easily [11]. In particular, the existence of large amount of continuous data may cause a challenging task to explicit concepts extraction from the raw data due to the huge amount of data space determined by continuous features [12]. So it is necessary to reduce the dimensionality of data and irrelevant factors before searching.

Image Retrieval
Feature selection is applied to content based image retrieval to allow efficient browsing, searching and retrieving [22]. Content based image retrieval is to index the images on the basis of their own visual contents (i.e. color, shape, texture etc.) instead of text based keyword indexing. The biggest problem for content based image retrieval is large a mount of images in database .

4. CONCLUSION

In this paper we have tried to provide an introduction to feature selection methods. The literature on feature selection methods is very vast. Many problems currently exist where the data encountered has many features, Feature selection may be employed to find which features will be useful so as computation may be focused on these. The desired effect of this is to speed up algorithms and also to make them more effective by only focusing on the relevant features in the data. Filter method is independent of classification algorithm and Wrapped and Embedded Method is dependent on classification algorithm.

5.REFERENCES

[1]. Anirban Dasgupta,Petros Drineas, Boulos Harb, Vanja Josifovski, Michael W. Mahoney; “Feature Selection Methods for Text Classification”,KDD- ACM.,2007.

[2]. Shyamala Doraisamy, Shahram Golzari, Noris Mohd. Norowi, Md. Nasir B Sulaiman, Nur Izura Udzir; ” A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music” Proceedings of ISMIR, 2008.

[3]. Gianluca Bontempi, Benjamin Haibe-Kains; “Feature selection methods for mining bioinformatics data”,

[4]. J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," in: C. Aggarwal (ed.), Data Classification: Algorithms and Applications. CRC Press, 2014.

[5]. N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method", *Expert Systems with Applications*, vol. 41, issue 14, pp. 6371–6385, 2014.

[6]. N.Gayatri, S.Nickolas, A.V.Reddy; “Feature Selection Using Decision Tree Induction in Class level Metrics Dataset for Software Defect Predictions”; Proceedings of the World Congress on Engineering and Computer Science ; Vol I; 2010.

[7].A jovik, K.Brick and N. Bogunovik “ A review of feature selection methods with application”.

[8]. Bo Pang, Lillian Lee; "Opinion Mining and Sentiment Analysis"; Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2 pp- 1–135, 2008.

[9]. Ms. Shweta srivastav,Ms. Nikita Joshi and Ms. Madhvi gaur "A review paper on feature selection methodology and their application"IJERD vol:7 2013.

[10]. Chih- Fong Tsai; "Data pre-processing by genetic algorithms for bankruptcy prediction", IEEE International Conference on Industrial Engineering and Engineering Management, Pages - 1780 - 1783 , 2011.

[11].Kyoung-jae Kim, Ingoo Han; "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index"; Expert Systems with Applications, Elsevier Science Ltd; 2000.

[12]. Liu, H., & Setiono, R.; "Dimensionality reduction via discretization"; Knowledge-Based Systems, 9 (1), 67–72; 1996.

[13]. Steppe. J., K.W. Bauer, "Feature Saliency Measures", Computers & Mathematics with Applications, Vol 33, No. 8, pp. 109-126; 1997.

[14]. T. Hastie, R. Tibshirani, and J. Friedman; "The Elements of Statistical Learning"; Springer, 2001.