International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

Survey on Hybrid Approach for Feature Selection

Aparna Choudhary[1], Jai Kumar Saraswat[2]

[1]Galgotias University, Greater Noida, India

[2]IA, Department of Information Technology & Communication, Jaipur, India

Abstract*: In text document categorization, feature selection (FS) is a strategy that aims at making text document classifiers more efficient and accurate. However, when dealing with a new task, it is still difficult to quickly select a suitable one from various FS methods provided by many previous studies. Feature selection, as a preprocessing step to machine learning, has been very effective in reducing dimensionality, removing irrelevant data, and noise from data to improving result comprehensibility. Researchers have introduced many feature selection algorithms with different selection criteria. However, it has been discovered that no single criterion is best for all applications. We proposed a hybrid approach for feature selection called based on genetic algorithms (GAs) that employs a target learning algorithm to evaluate features, a wrapper method. The advantages of this approach include the ability to accommodate multiple feature selection criteria and find small subsets of features that perform well for the target algorithm. In this way, heterogeneous documents are summarized and presented in a uniform manner.*

Keywords: Feature selection, Gene selection, Term selection, Dimension Reduction, Genetic algorithm, Text categorization, Text classification

1. Introduction

Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset.

Feature selection (known as subset selection) is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; one discards the remaining, unimportant dimensions. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction) [1]. There are two approaches:

- **Forward selection:** Start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error.
- **Backward selection:** Start with all the variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly); until any further removal increases the error significantly.

Two main models for feature selection are filtering and wrapper model [2]. The filtering approach receives a set of features, and filters it independently from the induction algorithm. The wrapper model searches for good feature subsets, and evaluates them using n-fold cross-validation on the training data. This scheme may be used in conjunction with any induction algorithm, which is used for evaluating feature subsets on the validation set. The search for feature subsets can be performed using simple greedy algorithms such as backward elimination or forward selection, or more complex ones that can both add and delete features at each step.

Since the wrapper model requires much more computation, filtering is the more common type of feature selection. This is especially true in the domain of textual information retrieval, where using the bag-of-words model results in a huge number of features. It was found that document frequency (DF), information gain (IG) and CHI are the most effective (reducing the feature set by 90-98% with no performance penalty, or even a small performance increase due to removal of noise). Contrary to a popular belief in information retrieval that common terms are less informative, document frequency, which prefers frequent terms (except for stop words), was found to be quite effective for text categorization.

1.1 Advantages of feature selection

It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed;

- It removes the redundant, irrelevant or noisy data.
- The immediate effects for data analysis tasks are speeding up the running time of the learning algorithms.
- Improving the data quality.
- Increasing the accuracy of the resulting model.
- Feature set reduction, to save resources in the next round of data collection or during utilization;
- Performance improvement, to gain in predictive accuracy;
- Data understanding, to gain knowledge about the process that generated the data or simply visualizes the data.

2. Related Works

Tao Liu [3] , Department of Information Science, Nankai University, Tianjin 300071, P. R. China, proposed an "Iterative Feature Selection (IF)" method that addresses the unavailability of label problem by utilizing effective supervised feature selection method to iteratively select features and perform clustering. Detailed experimental results on Web Directory data are provided in the paper.

Yaming Yang [4], School of Computer Science Carnegie Mellon University Pittusberg PA 15213-3702, USA , found

IG and CHI most effective in our experiments.Using IG thresholding with a k-nearest neighbor classifier on the Reuters corpus removal of up to 98% removal of unique terms actually yielded an improved classification accuracy.

Daniel I. Morariu, Lucian N. Vintan, and Volker Tresp [5] , World Academy of Science, Engineering and Technology 21 2006, presents three feature selection methods: Information Gain, Support Vector Machine feature selection called (SVM_FS) andGenetic Algorithm with SVM (called GA_SVM). We show that the best results were obtained with GA_SVM method for a relatively small dimension of the feature vector.

M. F. Zaiyadi and B. Baharudin [6] , World Academy of Science, Engineering and Technology 48 2010, proposed a novel hybrid approach for feature selection in text document categorization based on Ant Colony Optimization (ACO) and Information Gain (IG). We also presented state-of-the-art algorithms by several other researchers.

Cheng-Huei Yang and Li-Yeh Chuang [7] , proposed a filter method (information gain, IG) and a wrapper method (genetic algorithm, GA) for feature selection in microarray data sets. IG was used to select important feature subsets (genes) from all features in the gene expression data, and a GA was employed for actual feature selection. The K-nearest neighbor (K-NN) method with leave-one-out cross-validation (LOOCV) served as an evaluator of the IG-GA. The proposed method was applied and compared to eleven classification problems taken from the literature.

A.S. Kavitha, R. Kavitha, and J. Viji Grips [8] , introduced the classification accuracy using feature selection with machine learning algorithms. Feature selection is considered successful if the dimensionality of the data is reduced and accuracy of a learning algorithm improves or remains the same.

Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram [9], introduced two filters approaches namely Gain ratio and Correlation based feature selection have been used to illustrate the significance of feature subset selection for classifying Pima Indian diabetic database (PIDD). Genetic algorithm is used as search method with Correlation based feature selection as subset evaluating mechanism.

Isabelle Guyon, Andr´e Elisseeff [10] , Intoduced Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry.

Rajdev Tiwari, Manu Pratap Singh [11] , formulates and validates a method for selecting optimal attribute subset based on correlation using Genetic algorithm (GA), where GA is used as optimal search tool for selecting subset of attributes.

3. Future Enhancement

Some interesting research topics in feature selection of potential impact in the near future. The proliferation of feature selection techniques brought out the difficulty in choosing the best suitable feature selection algorithm for an application, which is resulted from the different feature selection criteria employed by different feature selection algorithms. We proposed a **hybrid algorithm for feature selection** to solve the problem. Our objective is implementation of hybrid algorithm on datasets for identifying the crucial feature subset that is capable of generating accurate predictions.

4. Conclusion

In this survey we had projected various feature selection methods, terms, limitations, advantages and available recent innovation in feature selection. We hope, that the interested readers will have broad overview of this field and several starting point for further details. Feature selection remains and will continue to be an active field that is incessantly rejuvenating itself to answer new challenges.

References

[1] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol.34, No. 1, pp. 1–47, 2002.
[2] YZhao and G Karypis, "Hierarchical clustering algorithms for document datasets", Data Mining and Knowledge Discovery, pp. 141-168, 2005.
[3] Tao Liu, "An Evaluation on Feature Selection for Text Clustering", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
[4] Yaming Yang, "A comparative study on Feature Selection in Text Categorization", School of Computer Science Carnegie Mellon University Pittusberg, PA 15213-3702, USA
[5] Daniel I. Morariu, Lucian N. Vintan, and Volker Tresp, " Evolutionary Feature Selection for Text Documents using the SVM ", World Academy of Science, Engineering and Technology 21 2006
[6] M. F. Zaiyadi and B. Baharudin," A Proposed Hybrid Approach for Feature Selection in Text Document Categorization", World Academy of Science, Engineering and Technology 48 2010
[7] Cheng-Huei Yang and Li-Yeh Chuang," IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data ", J. Med. Biol. Eng., Vol. 30. No. 1 2010
[8] A.S. Kavitha, R. Kavitha, and J. Viji Grips," Empirical Evaluation of Feature Selection Technique in Educational Data Mining ", VOL. 2, NO. 11, Dec 2012 ISSN 2225-7217, ARPN Journal of Science and Technology ©2011-2012. All rights reserved. http://www.ejournalofscience.org
[9] Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram," Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection ", International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 271-277
[10] Isabelle Guyon, Andr´e Elisseeff," An Introduction to Variable and Feature Selection ", Journal of Machine Learning Research 3 (2003) 1157-1182
[11] Rajdev Tiwari, Manu Pratap Singh," Correlation-based Attribute Selection using Genetic Algorithm ", International Journal of Computer Applications (0975 – 8887)Volume 4– No.8, August 2010