Feature Selection Techniques for Text Classification

A. Mohammad Behrouzian Nejad, B. Sayed Mohsen Hashemi, C. Aref Sayahi and D. Behnam Kiaeimehr

Sama Technical and Vocational Training College, Islamic Azad University, Shoushtar Branch, Shoushtar, Iran
Department of Computer Engineering, Soosangerd Branch, Islamic Azad University, Soosangerd, Iran
Department of Computer Engineering, Soosangerd Branch, Islamic Azad University, Soosangerd, Iran
Sama Technical and Vocational Training College, Islamic Azad University, Shoushtar Branch, Shoushtar, Iran
Mohamad.behrouzian@gmail.com

Abstract

Text classification of documents refers to classifying documents to one or more predefined classes. One of the most important steps in text classification is feature selection. In text classification, feature selection is a strategy that can be used to increase the efficiency and accuracy of classification. Feature selection techniques can be classified into two basic categories: filtering techniques and wrapper techniques. Filtering techniques are independent of the learning algorithm. But wrapper methods uses from learning algorithm as the evaluation function. In this paper we review some effectiveness feature selection researches and show review results of these in a table form.

Keywords: Text Classification, Feature Selection, Filtering, Wrapper.

I. Introduction

Nowadays significant parts of the information are stored in textual databases (or text documents) which are composed of a large set of documents from various sources, such as news, articles, books, digital libraries, e-mail messages and Web pages. Using Automatic Text Classification methods can be classifying this data. Text classification of documents means that the documents which not classified, we assign into one or more predefined categories. One of the most important steps in text classification is feature selection phase (Dave, 2011; Nejad *et al.*, 2013). Feature selection process refers to choose a subset of features of the text (words). According to this problem which features in text documents are very much and this affects to reduce of classification performance; in feature selection step, basic goal is select the best features that help to classification performance. By removing irrelevant features and non-discrimination, classification performance can be increased (Sebastiani, 2002; Jensen, 2005). In this paper we discuss feature selection in text classification and review some effective methods that presented for text classification.

II. Feature Selection

Feature selection refers to the selection of those features that are more important. As many systems are large scale in various areas of data collection, feature selection is an important and widely grown. Some of the basic applications of features selection have been shown in fig. 1(Jensen, 2005).

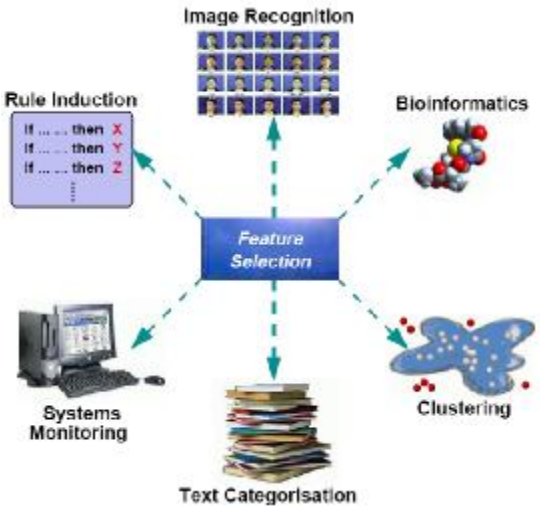

Figure. 1. Applications of features selection

Most feature selection algorithms are used to optimize the performance of classification systems for visual recognition. For clustering, categorization of text documents as a set of words has many applications. Documents with keywords extracted component and according to criteria such as the occurrence of repeated words are examined. Due to the fact that the numbers of extracted keywords usually are very much, reducing feature sizes should be done. This can be done using simple filtering methods such as words stemming or delete stop words and can reduce the dimensions of features. However, this reduction is not sufficient for use in automatic classification, so a feature selection method should be used.

Feature selection techniques can be classified into two basic categories: filtering techniques and wrapper techniques. Filtering methods are independent of the learning algorithm. These methods Regardless of learning algorithm and using statistical methods to feature selection and have low complexity. But wrapper methods uses from learning algorithm as the evaluation function. These methods have higher time complexity and accuracy than filter methods (Nejad *et al.*, 2013). With the increasing size of the features in text classification, generally these methods could not be used because of the high complexity. Some filtering methods that can be used in many texts classification techniques such as Document Frequency (DF), Information Gain (IG), and Mutual Information (MI). Some wrapper methods are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS) and Neural Networks (Dave, 2011; Eyheramendy and Madigan, 2005).

III. Related Researches

In text classification, feature selection is a strategy that can be used to increase the efficiency and accuracy of classification. Authors in (Bryll *et al.*, 2003) proposed Attribute Bagging (AB) that is a technique to improve the stability and overall accuracy of classification using a random subset of features. AB is a wrapper method that can be applied to any learning algorithm. This method can create subsets of features with suitable size and then randomly select subsets of features to make the classification scheme of the training set on the hands. Combining classifiers can be used for voting. This paper evaluates proposed technique with using bagging and other algorithms on a hand pose data set. Result show that this method has better performance than bagging in stability and accuracy criteria.

Authors in (Li *et al.*, 2009) presented theoretical framework for feature selection methods are based on two principles: frequency measurement and ratio measurement. Then, evaluate six feature selection methods within this framework. The authors proposed method called frequency weighted odds (WFO) that two principal measurement is combined with the trained weight. Evaluation results on two topic-based and sentiment classification tasks datasets show that proposed method have good performance in different tasks and selected features.

In (Li and Hsu, 2009) from a combination of five filter feature selection methods: document frequency, information gain, chi-square, mutual information and the term strength has been used. After combine obtained feature vectors are considered as the best features. In combination step used from rank and score combination. After determining the new features, the features with highest score are identified as the optimal features. Results show that rank combination has better performance than score combination.

Authors in (Ogura *et al.*, 2009) proposed a new feature selection method. The authors have used the degree of deviation of a Poisson distribution for the selection of useful features. Authors in (Aghdam *et al.*, 2009) for increasing classification performance used from Ant Colony Optimization (ACO). ACO method use from real actions of ants to finding shortest path to food. For feature selection in this method, a complete graph of features is drawn first, and then optimal subset is extracted. Proposed method evaluated with genetic algorithm, information gain and chi-square on reuters-21578 dataset. Results show that proposed method has good performance.

Uguz in (Uguz, 2011) to reduce the complexity use from benefits of two filtering and wrapper feature selection methods. Due to the low complexity of filter techniques, initially selected feature using information gain and bulk properties have to remove less important features. Then in the next step, the due to better results of wrapper techniques than filtering techniques, from extracted features by using information gain, two wrapper methods of genetic algorithm and principal component analysis (PCA) algorithms used for the find best subset of features. Evaluation results show that combining genetic algorithm and information gain have better performance than information gain. Combining PCA and information gain have similar or weaker performance than information gain.

According to this note that the feature selection metrics generally are based on the document frequency or term frequency, authors in (Azam and Yao, 2012) do a comparison between feature selection metrics based on term frequency and document frequency in text classification. In this evaluation provides important information which term frequency is used in the feature selection function. For this, two gini index and Discriminative Power Measure (DPM) methods are used. Evaluations run on the reuters-21578 dataset and results show that metrics based on term frequency have better performance in low features. Metrics based on document frequency have better performance in more features.

IV. Conclusion

In text classification, feature selection is a strategy that can be used to increase the efficiency and accuracy of classification. According to this note that wrapper feature selection methods have high complexity, we propose use from filtering methods to select optimal subset of features in text classification because in text, number of feature is high. Also we propose that use from hybrid techniques instead single techniques to increase performance of feature selection. In this paper we review some effectiveness feature selection researches that table I show the characteristics these methods.

TABLE I
CHARACTERISTICS FEATURE SELECTION METHODS

References	Characteristics
Bryll *et al.*, 2003	• Improve stability and accuracy of classification • Select subset of features
Li *et al.*, 2009	• Use theoretical framework for feature selection based on frequency measurement and ratio measurement • Efficiency in different number of features
Li and Hsu, 2009	• Use of combined feature vector to select the best features • Use from rank and score combination • Better performance of rank combination than score combination
Aghdam *et al.*, 2009	• Feature selection by ACO concept • Create complete graph of features and select the best subset of features • Better performance than genetic algorithm and information gain
Uguz, 2011	• Use of benefits of filtering and wrapper methods • Use of genetic algorithm and PCA together information gain to select best subset of features • Better performance of genetic algorithm and information gain combination
Azam and Yao, 2012	• Comparing feature selection metrics based on document frequency and term frequency • Increase efficiency in low number of features for metrics based on term frequency • Increase efficiency in high number of features for metrics based on document frequency

References

i. Aghdam MH, Aghaee NG, Basiri ME (2009). Text feature selection using ant colony optimization, Expert Systems with Applications, 36(3): 6843-6853.
ii. Azam N, Yao JT (2012). Comparison of Term Frequency and Document Frequency Based Feature Selection Metrics in Text Categorization, Expert Systems with Applications, 39(5): 4760-4768.
iii. Bryll R, Osuna RG, Quek F (2003). Attribute Bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recognition, 36: 1291-1302.
iv. Dave K (2011). Study of feature selection algorithms for text-categorization, University of Nevada, Las Vegas, UNLV Theses/Dissertations/Professional Papers/ Capstones, Paper 1380.
v. Eyheramendy S, Madigan D (2005). A novel feature selection score for text categorization, In Proceddings of the International Workshop on Feature Selection for Data Mining, Newport Beach, CA.
vi. Jensen R (2005). Combining rough and fuzzy sets for feature selection, PhD Thesis, University of Edinburgh, UK.
vii. Li S, Xia R, Zong C, Huang CR (2009). A Framework of Feature Selection Methods for Text Categorization, In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP: 692-700.
viii. Li Y, Hsu DF, Chung SM (2009). Combining Multiple Feature Selection Methods for Text Categorization by Using Rank-Score Characteristics, In Proceedings of the 21st International Conference on Tools with Artificial Intelligence, New Jersey, USA: 508-517.
ix. Nejad MB, Attarzadeh I, Hosseinzadeh M (2013). An Efficient Method for Automatic Text Categorization", International Journal of Mechatronics, Electrical and Computer Technology, 3(9): 314-329.
x. Ogura H, Amano H, Kondo M (2009). Feature selection with a measure of deviations from Poisson in text categorization, Expert Systems with Applications, 36(3):6826–6832.
xi. Sebastiani F (2002). Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1): 1-47.
xii. Uguz H (2011). A Two-Stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm, Knowledge-Based Systems, 24: 1024-1032.