A SURVEY ON RELEVANCE FEATURE SELECTION METHOD FOR TEXT CLASSIFICATION

Nisha Ranjani.S[1]
PG Scholar,
Department of cse,
SNS College of Engineering
Coimbatore.

Karthikeyan.K[2]
Assistant Professor,
Department of cse,
SNS College of Engineering,
Coimbatore.

ABSTRACT

To guarantee the quality of relevance feature in text documents is a big challenge because of large scale terms and data patterns. All the existing approaches have suffered from the two major problems such as polysemy and synonmy .There are several hybrid approaches were proposed for text classification. Feature selection is the process of selecting a subset of feature used to represent the data. In text classification it focuses on identifying relevant information without affecting the accuracy of the classifier. This paper gives the surveys on several approaches of text classification and feature selection methods for text classification. Feature selection methods are discussed for reducing the dimensionality of the dataset by removing features that are irrelevant for the classification.

Key words: Feature selection; Text classification; Relevance feature.

I INTRODUCTION

Text mining has a large range of applications such as text summarization, categorization, entity and sentimental analysis. Text mining requires pre-processing which the text must be decomposed into smaller units such as terms and phrases. For example, in some text mining applications, terms extracted from the documents and treated as features. Text clustering is also termed as document clustering. Clustering is used to group the entire documents into relevant topics. Each of that group is known as clusters. This is an unsupervised learning technique. The major problem in document clustering is its high dimensionality. It requires efficient algorithms which can solve this high dimensionality clustering. Several algorithms are used for text clustering which includes partitioning clustering algorithm, Density-based clustering algorithm, Model-based clustering algorithm, Hierarchical clustering algorithm and frequent pattern-based clustering. The high dimensionality of data is the great challenge for effective text categorization is high dimensionality. Each document in a document quantity contains much noisy and irrelevant information which may reduces the efficiency for text categorization. Most of the categorization techniques reduce this features by eliminating stemming or stop words. It is important to use feature selection methods to hold the high dimensionality of data for effective text categorization. In text classification Feature selection mainly focuses on identifying relevant information without affecting the accuracy of the classifier. The objective of the feature selection to find the useful patterns in the text documents. Feature reduction will transform the original features into new features by applying some transformation function. This new feature set contains fewer features or dimensions than the original set. It yields good results over a reduced dimension feature space. The challenging problem is how to use the patterns to weight accurately.

II CLASSIFICATION METHODS

There classification is used to classify the datasets as two ways such as

- Supervised feature selection and
- Unsupervised feature selection.

A. Feature selection methods:

i) Information Gain (IG)

Information gain is a popular feature selection method in text categorization .It measures the number of bits of information obtained for a particular category by the presence of the term in a document or absence of the term. IG score will be null for the two independent variables. It will also be high because of the dependency between two variables. IG feature selection method selects the terms having the highest information gain scores. The information gain for a term t is defined as

$$IG(T) = \sum_{i=1}^{m} X(Ci) log X(Ci) + X(t) \sum_{i=0}^{m} X\left(\frac{Ci}{t}\right) log X\left(\frac{Ci}{t}\right) \quad (1)$$

Where m is the number of classes

ii) X^2 statistics

Chi-square statistics is frequently used in feature selection to measure how the observed results are differing from the expected results. In other words, it will measure the independence between two random variables. The two random variables in text categorization are occurrence of term t_k and class c_i. The chi-square statistics measures the independence between t_k and c_i. The formula for measuring chi-square score is:

$$CHI(t_k, c_i) = N_x \frac{[P(t_k, c_i) P(\overline{t_k}, \overline{c_i}) p\ (t_k, \overline{c_i}]^2}{P(t_k) P(\overline{t_k}) P(\overline{c_i})} \quad (2)$$

The chi-square method selects the terms with the highest chi-square score which is more useful for classification.

iii) Document Frequency (DF)

Document frequency is the measure of number of documents in which a single term occurs in a dataset. It is the simplest principle for term selection and it easily scales in to a large dataset. It is simple method but an effective feature selection for the text categorization. This method is to eliminate the rare words which are assumed confusing for classification. Document frequency method selects the terms with the scores. Its formula is

$$DF(t_k, c_i) = P(t_k, c_i) \quad (3)$$

B. Classification Methods

i) K-Nearest Neighbor

KNN is a case based learning algorithm. The objects are classified by selecting several labeled terms with their smallest distance from each object. The Major disadvantage of KNN is that it uses all features in computing distance and costs very much time for classifying objects. The classification is usually performed by comparing the class frequencies of the k nearest documents. The evaluation is done by measuring of angle between the two feature vectors. Feature vectors have to be normalized to length 1. The main advantage of the k-nearest neighbor method is its simplicity. Its disadvantage is that it requires more time for classifying objects when there is a large number of training examples.

ii) Decision Trees

Decision tree methods will reconstruct the manual classification of the documents by constructing well-defined queries (true/false) in the form of a tree structure where the nodes represent the questions and the leaves represent their corresponding category of the documents. After the tree is created, a new document can be easily be classified by putting them in to root node of the tree and run through their query structure until certain leaf is reached. The advantage of decision trees is that the output tree is easy to understand even for persons who are not familiar with the details of the model. The structure of tree is generated by the model which provides the user with combined view of the classification logic. A hazard of the application of tree methods is "over fitting" that is if a tree over fits then training data will classifies the training data bad but it would classify the documents to be categorized later better.

iii) Naïve Bayesian Approaches

In Bayesian approaches there are two groups in document categorization:

- Naïve and

- Non-naïve Bayesian.

The naïve part of the previous is the assumption of word (i.e. feature) independence, that the word order is irrelevant and the presence of one word will not affect the presence or absence of another word. The disadvantage of Bayesian approaches is that it can only process binary feature vectors.

iv) Neural network

It is a network of units, where the inputs and are usually represent terms and Category. For classifying a document, their term weights are assigned to the input units, the unit activation is propagated through the network, and then the output unit(s) takes up as a outcome determines the categorization. Due to its simplicity of implementation researches uses the single-layer percepton. The multi-layer perceptron which is more complicated, also usually implemented for classification tasks. Models using two types of network for document classification

- back-propagation neural network(BPNN) and
- modified back-propagation neural network (MBPNN)

v) Support Vector-based Methods

There are two types of vector-based methods:

- Centroid algorithm and
- Support vector machines.

One of the simplest methods is the centroid algorithm. During the learning stage average feature vector for each category is calculated and it will be set as centroid-vector for the each category. A document is easily categorized by discovering the centroid-vector closest to its feature vector. The method is also improper when the number of categories is very large. Support vector machines (SVM) need positive training documents and also a certain number of negative training documents. SVM is looking for the decision surface that separates the positive term from the negative examples in the n-dimensional space. The document is closest to the decision surface are called support vectors. The algorithm results remain unchanged if documents that not belongs to the support vectors and they are removed from the training dataset. An advantage of SVM is runtime-behavior during the categorization of new documents. A disadvantage is that a document is assigned to various categories because of the similarity calculated individually for each category.

III. APPLICATIONS OF TEXT CLASSIFICATION

The applications of text categorization are manifold. Common traits among all of them are

- The need to handle and organize documents in which the textual component is either unique, or dominant, or simplest to interpret component.
- The need to handle and organize large quantities of such documents, large enough that their manual organization into classes is either too expensive or not feasible within the time constraints imposed by the application.
- The fact that the set of categories is known in advance, and is variation over time is small.

IV.CONCLUSION

In this survey paper we discuss the types of feature selection method in text mining for discovering the relevance features. Many algorithms were discussed with their issues with advantages and disadvantages. Our future work is to find the efficient algorithm to overcome the issues in the existing algorithms. Many approaches for text categorization are discussed in this paper. Feature selection methods are able to successfully reduce the problem of dimensionality in text categorization applications. Process of text classification is well researched, but still many improvements can be made both to the feature preparation and to the classification engine itself to optimize the classification performance for a specific application. Research describing what adjustments should be made in specific situations is common, but a more generic framework is lacking. Effects of specific adjustments are also not well researched outside the original area of application. Due to these reasons, design of text classification systems is still more of an art than exact science.

V.REFERENCES

[1] Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan, *A Survey on Data Mining Techniques*, International Journal of Advanced Research in Computer and Communication Engineering, Volume 3, Issue 1, 2014, pp. 5002-5003.

[2] Daniel Engel, Lars Hüttenberger, Bernd Hamann, A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization, LNCS Springer, 2014, pp. 1-16.

[3] Khalid, Samina, Khalil Tehmina, Nasreen Shamila, A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning, IEEE Science and Information Conference, 2014, pp. 372-378.

[4] Azam.N and Yao.J, “Comparison of term frequency and document frequency based feature selection metrics in text categorization,” Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.

[5] Chandrashekar .G and Sahin.F, “Asurvey on feature selection methods,” in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[6] Li.Y, Hus.D.F, and ChungS.M, “Combination of multiple feature selection methods for text categorization by using combinational fusion analysis and rank-score characteristic,” Int. J. Artif. Intell. Tools, vol. 22, no. 2, p. 1350001, 2013.

[7] Salton.G and Buckley.C, “Term-weighting approaches in automatic text retrieval,” in Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Aug. 1988.

[8] Song.Q, Ni.J, and Wang.G, “A fast clustering-based feature subset selection algorithm for high-dimensional data,” in IEEE Trans.Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.

[9] Yang.Y and , Pedersen.J.O, “A comparative study on feature selection in text categorization,” in Proc. Annu. Int. Conf. Mach.Learn., 1997, pp. 412–420.