Research Paper

A NOVEL METHOD FOR TEXT CLASSIFICATION USING K-MEANS CLUSTERING AS FEATURE SELECTION

SURBHI JAIN¹, SARVESH SINGH RAI²

². Asst. Prof., Dept. of Computer Science & Engineering, Infinity Management & Engineering College, Sagar (M.P.).

ABSTRACT

Text mining has been employed in a wide range of applications such as text summarization, text categorization, named entity extraction, and opinion and sentimental analysis. Text classification is the task of assigning predefined categories to free-text documents. Clustering of documents is used to group documents into relevant topics. Each of such group is known as clusters. It is an unsupervised learning technique. The major difficulty in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. The high dimensionality of data is a great challenge for effective text categorization. In this paper, we discuss a text categorization method based on k-means clustering feature selection. K-means is classical algorithm for data clustering in text mining, but it is seldom used for feature selection. For text data, the words that can express correct semantic in a class are usually good features. We use k-means method to capture several cluster centroids for each class, and then choose the high frequency words in centroids as the text features for categorization. The words extracted by k-means not only can represent each class clustering well, but also own high quality for semantic expression. On normal text databases, Regularized least-squares regression based on our feature selection method exhibit better performances than original classifiers for text categorization.

INDEX TERMS

Feature selection, k-mean clustering, feature clustering and Regularized least-squares regression.
INTRODUCTION

Text mining has been employed in a wide range of applications such as text summarization, text categorization, named entity extraction, and opinion and sentimental analysis. Text mining requires a great deal of pre-processing in which the text must be decomposed into smaller syntactical units (e.g., terms and phrases). Sometimes, the text data may also need to be transformed into other types. For example, in some text mining applications, terms extracted from the documents in the entire corpus are treated as features and documents are treated as records. Thus, each document can be represented as a Boolean vector in which a true (or false) value for a feature indicates the presence (or absence) of the corresponding term in the document. During mining stage, depending on the requirements of the specific applications, various data mining methods such as association mining, classification and clustering are frequently used in text categorization applications.

Generally, text-mining refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text classification is the task of assigning predefined categories to free-text documents. That is, it is a supervised learning technique. While in text clustering the possible categories are unknown and need to be identified by grouping the texts. Text clustering is also called document clustering. Clustering of documents is used to group documents into relevant topics. Each of such group is known as clusters. It is an unsupervised learning technique. The major difficulty in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. There are several algorithms for text clustering which includes agglomerative clustering algorithm, partitioning clustering algorithm, Hierarchical clustering algorithm, Density-based clustering algorithm, Grid-based clustering algorithm, Model-based clustering algorithm, frequent pattern-based clustering, and Constraint-based clustering.

The high dimensionality of data is a great challenge for effective text categorization.
Each document in a document corpus contains much noisy and irrelevant information which may reduce the efficiency of text categorization. Most text categorization techniques reduce this large number of features by eliminating stemming or stop words. It is effective to a certain extent but the remaining number of features is still huge. It is important to use feature selection methods to handle the high dimensionality of data for effective text categorization. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. More sophisticated feature selection techniques have been reported in the literature [1, 2, 3].

Text classification is commonly used to handle spam emails, classify large text collections into topical categories, used to manage knowledge and also to help Internet search engines. A major characteristic of text categorization is high dimensionality of the feature space. The native feature space consists of hundreds of thousands of terms for even a moderate sized text collection.

Email classification is considered as supervised classification problem, and also categorized data are easy for user to browse. In a supervised setting, given a supervision in the form of a set of labeled training examples (e.g. e-mails labeled as belonging to different folders such as spam, legitimate, work, teaching etc.), the goal is to build a classifier, that is then used to predict the category of an unseen incoming e-mail.

FEATURE SELECTION METHODS

The main plan of Feature selection (FS) is to rank terms according to how well they differentiate between object categories. The output of the feature selection method could be a ranking of features supported the applied feature selection algorithm [43].

1) Information Gain (IG)  Information gain is used for ranking of features based on finding how well they differentiate classes like spam, legitimate, phishing etc., Information gain work similar as used for spitting criteria for decision tree. The entropy $I$ of a given dataset $S$ is outlined as –

$$I (S) = -\sum_{i=1}^{C} p_i \log_2 p_i \quad (1)$$
Where $C$ denotes the total number of classes and $p_i$ the portion of instances or feature that belong to class $i$. The reduction in entropy or the information gain is computed for every attribute $A$ consistent with

$$IG(S, A) = I(S) \sum_{v \in A} \frac{|S_{A,v}|}{|S|} I(S_{A,v}, V)$$

(2)

Where $v$ is a value of $A$ and $S_{A,v}$ is the set of instances where $A$ has value $v$ [4].

2) Document Frequency (DF)

Document frequency is a very simple feature selection method. Document frequency for a term can be calculated by counting number of documents in which a term/feature occurs. In line with the DF methodology, the terms whose DF is below a predefined threshold are removed from the set of terms. The DF of a term is outlined as follows:

$$D(t_i) = | \{m_j \mid m_j \in M, \text{ and } t_i \in m_j \} |$$

(3)

Where $M$ denotes the complete training set of messages, and $m_j$ is a message in $M$. The essence of DF is to get rid of rare terms. In step with its assumption, rare terms offer very little information for classification, therefore the elimination of them doesn't have an effect on overall performance. As shown in [5], the performance of DF was similar to that of information gain and statistic (CHI) with up to will increase term elimination. One major advantage of document frequency is that its complexity quality increases linearly with computational of training messages [6].

3) Term Frequency Variance

Koprinska et al. [46] developed a TFV technique to pick out the terms with high variance that was thought-about to be additional informative. Like information gain term frequency variance method is category dependent. For each term $f$ we compute $w$ the term/document frequency (tf) in every class and variance can be calculated as

$$TFV(f) = \sum_{i=1}^{k}[tf(f, c_i) - \text{mean}_tf(f)]^2$$

(4)

TFV is seen as an improvement of the document frequency that is the simplest
technique for feature reduction. For every term within the training corpora, document frequency counts the number of documents during which the term occurs and selects features with frequency higher than a predefined threshold. The idea is that rare terms don't seem to be informative for class prediction.

4) Mutual Information

Mutual information may be a criteria usually employed in statistical language modeling of word association and connected application. If one consider the manner contingency table of a term t and class c, wherever A is number of time t and c co-occur, B is number of time the t occur while not c, C is range of time c occur while not t, and N is the total number of documents, then the mutual information criteria between t and c is outlined to be

$$I(t, c) = \log \frac{\Pr(t \cap c)}{\Pr(t) \cdot \Pr(c)} \quad (5)$$

And is estimated using

$$I(t, c) = \log \frac{A + N}{(A + C) \cdot (A + B)} \quad (6)$$

I (t, c) has a natural value of zero if t and c are independent [47].

5) $\chi^2$ statistic (CHI)

The $\chi^2$ statistic measures the shortage of independence between t and c and might compared to the $\chi^2$ distribution with one degree of freedom to judge extremeness. exploitation the two method contingency table of a term t and class c, wherever A is that the number of time t and c co-occur, B is number of time the t occur while not c, C is number of time c occur while not t, D is that the number of times neither c or t occur, and N is the total number of documents, the term goodness measure is outline to be:

$$\chi^2 = \frac{N \cdot (A D - C B)^2}{(A + C) \cdot (B + D) \cdot (A + B) \cdot (C + D)} \quad (7)$$

The $\chi^2$ statistic has a natural value of zero if t and c are independent [7].

occur while not t, D is that the number of times

PROPOSED METHOD

In this paper we have perform experiments on 20 newspaper dataset. We collected dataset from UCI Machine Learning
Repository. First task is to collect term from the dataset then apply a feature selection process for reducing the dimension of dataset, for that purpose we have used k-mean clustering algorithm for finding best feature from the dataset. These feature is forwarded to next step of classification. For classification we have used Regularized least squares (RLS) using lasso or elastic net algorithms.

4.1. K-means

K-means is one of the simplest clustering algorithms to group data, which aims to partition the samples into k sets with minimizing cluster error. In k-means there are three main steps, first selecting k initial cluster centroids, second assigning each sample to the nearest centroid, and final updating the centroids by the means for each cluster. We briefly give the process of k-means:

1. Initial cluster centroids (m1, m2...mk) are randomly selected from given samples (x1, x2...xn).

2. The similarities between each sample and all centroids are computed, and then each sample is assigned to the nearest centroid.

3. The means of samples in each cluster are calculated as the new cluster centroids.

The step (2) and (3) are repeated until the final stable clustering results are obtained.

For text data, the similarity between a sample and a centroid in k-means usually adopts Euclidean distance and Cosine distance. In the following, we give the two distances.

1. Euclidean distance

\[ D(x_i, m_j) = \sqrt{\sum_{i=1}^{n} (x_i - m_j)^2} \]

(x1,x2,...xn) are samples, and (m1,m2,...mk) are the clustering centroids. The distance between sample and centroid adopted in k-means directly affects the clustering results, and the final centroids will have the minimal means of distances.

4.2. Regularized least squares (RLS)

Regularized least squares (RLS) is a family of methods for solving the least-squares problem while using regularization to further constrain the resulting solution.

LASSO (Least Absolute Shrinkage and Selection Operator)
For a given value of \( \lambda \), a nonnegative parameter, lasso solves the problem

\[
\min_{\beta_0, \beta} \left( \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),
\]

Where
1. \( N \) is the number of observations.
2. \( y_i \) is the response at observation \( i \).
3. \( x_i \) is data, a vector of \( p \) values at observation \( i \).
4. \( \lambda \) is a nonnegative regularization parameter corresponding to one value of Lambda.
5. The parameters \( \beta_0 \) and \( \beta \) are scalar and \( p \)-vector respectively.

As \( \lambda \) increases, the number of nonzero components of \( \beta \) decreases.

The lasso problem involves the L1 norm of \( \beta \), as contrasted with the elastic net algorithm.

**Elastic Net**

For an \( \alpha \) strictly between 0 and 1, and a nonnegative \( \lambda \), elastic net solves the problem

\[
\min_{\beta_0, \beta} \left( \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta) \right),
\]

where

\[
P_\alpha(\beta) = \frac{(1-\alpha)}{2} \| \beta \|_2^2 + \alpha \| \beta \|_1 = \sum_{j=1}^{p} \left( \frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j| \right).
\]

Elastic net is the same as lasso when \( \alpha = 1 \). As \( \alpha \) shrinks toward 0, elastic net approaches ridge regression. For other values of \( \alpha \), the penalty term \( P_\alpha(\beta) \) interpolates between the L1 norm of \( \beta \) and the squared L2 norm of \( \beta \) [27-30].

**EXPERIMENTS**

In this paper, based on k-means feature selection we will discuss two classifiers, Regularized least squares (RLS) and SVM method for text categorization. The experiments are conducted on text corpus, DBWorld6. We split each corpus into two parts as training and test set. On training set, we conduct our algorithm to select features for text expression, and then represent all the samples by the selected features. After the feature selection, we run two classifiers, RLS and SVM. We compared the methods based on the selected features with original classifiers. The normal accuracies of text categorization, macro-F score and micro-F score, and the running time of classification are tested. In Table 1, we give the comparison results.
Table 1. Experimental Results on DB World dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Macro-F (%)</th>
<th>Micro-F (%)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>80.25</td>
<td>81.25</td>
<td>9.56</td>
</tr>
<tr>
<td>RL</td>
<td>81.3</td>
<td>84.8</td>
<td>9.00</td>
</tr>
</tbody>
</table>

CONCLUSION

Text document contains a large number of features. Therefore appropriate and accurate feature selection techniques are generally essential to the performance of text classification systems. In a text document, each word can be a possible feature.

In this paper, we use k-means clustering method to collect and choose features for text categorization. As the words in clustering centroids of each class can represent class well, thus we choose the features with larger word-frequency for text categorization. Experiments on DB World text corpus show that the capacities of text classifiers will be enhanced by k-means feature selection using Regularized least squares method.

REFERENCES

The use of MMR, diversity-based reranking for reordering documents and producing summaries.


27. Friedman, J., R. Tibshirani, and T. Hastie. Regularization paths for
http://www.jstatsoft.org/v33/i01