Comparison of Feature Selection and Classification Algorithms for Restaurant Dataset Classification

KOMKID CHATCHARAPORN¹, NARODOM KITTIDACHANUPAP¹, KITTISAK KERDPRASOP², NITTAYA KERDPRASOP²

¹School of Information Technology
²Data Engineering Research Unit, School of Computer Engineering
1,2Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000, THAILAND
komkid1@hotmail.com, narodom_k@hotmail.com, kerdpras@sut.ac.th, nittaya@sut.ac.th

Abstract: - Currently, the rapid growth of information on the Internet makes automatic text classification play an important role to help people discovering desired information on enormous resources. Text mining, feature selection and classification algorithm have effect on the classification performance directly. In this paper, the comparative study of the text classification performance is proposed. It compares between three feature selection approaches with four classification algorithms. The algorithms include C4.5, ID3, Bayes Net and Naïve Bayes. The objective is to find suitable model for restaurant dataset. The experimental results indicated that when a set of selected features was large, C4.5 and ID3 algorithms generated models with better accuracy than those induced by Bayes Net and Naïve Bayes algorithms. On the other hand, the small amount of selected features makes every algorithm yields models with almost the same accuracy. However, algorithms based on Bayes’ theorem could generally induce classification models with higher accuracy than the tree-based algorithms.

Key-Words: - Classification Algorithms, Feature Selection, Restaurant Classification, Text Classification

1 Introduction

The number of online texts on the Internet is increasing rapidly. This makes data mining an important technology to deal with such large amount of data. The main benefit of data mining is to help people to extract and discover valuable information from tremendous resources. There are many data mining techniques that have been used in business such as classification, clustering, and association analysis.

Text classification is one type of classification tasks. It can automatically assign natural language texts based on their content to predefined classes or categories [1]. The demands of text classification are also increased in present time. The information overload problem and the semi-structured contents of most texts are two major causes of unsuccessful classification by humans. Hence, machines can play important role to cope with these problems by their capability to perform automatic text classification.

Text classification can be applied to classify online contents such as news, comments, products, books, movies as well as travel attractions [2]. In these works, there are still some considerable issues for text classification. The first issue is data sources; many sites on the Internet still have many uncategorized contents. This problem can be solved with text classification. Yet, gathering contents on those sites to create classification models should be considered about the quality of data [3]. The second issue is the overfitting. The overfitting is drawback of classification. It can make classification model always provide very high accuracy with training dataset but it cannot provide good accuracy with test dataset [4]. These issues reflect the problems of text classification.

The feature selection and the performance of classification algorithms are important issues for text classification, because they have directly impact on the quality of classification model [5]. The proper feature selection method can identify significant features that have related with each class [6]. The appropriate selection of classification algorithm can produce model with good predictive performance. Furthermore, the collected information for creating the classification model should be suitable and adequate in details [7].

In this paper, the comparative study of the text classification’s performance is proposed. It compares between three different feature selection
approaches and four classification algorithms. The algorithms include C4.5, ID3, Bayes Net, and Naïve Bayes [7]. The purpose of the comparison is to find suitable model for classifying restaurant dataset. The datasets only focuses on restaurant datasets because restaurants are one kind of popular places of tourist attraction. The training and test restaurant datasets are gathered from the Lonely Planet.com website.

2 Feature Selection and Related Work

2.1 Feature Selection Methods

The major problem of text classification is the curse of dimensionality when it has too many words in feature set [8]. The feature set in text classification is a set of unique terms or words that appear in all documents. Feature selection is a method that has been used to reduce the amount of features by selecting the subset of relevant features from original set. Furthermore, the method can improve performance of classification and avoid the overfitting problem. In most text classification tasks, the six feature selection methods are normally applied: Term Frequency, Inverse Document Frequency (TF-IDF), Information Gain (IG), Gain Ratio (GR), Chi Square, ReliefF and Support Vector Machine (SVM).

2.1.1 Term Frequency, Inverse Document Frequency (TF-IDF): TF-IDF is weighing method which is used in information retrieval and text classification. It uses statistical measure to evaluate importance of a word in documents. The high weight value indicates that the word has high ability to separate document [9]. The weight calculation is shown in equation (1).

\[ W_{(f,d)} = TF_{(f,d)} \times \log \frac{|D|}{|DF_{(f)}|} \]  

where 
- \( W_{(f,d)} \) is weight of a feature \( f \) that occurs in a document \( d \),
- \( TF_{(f,d)} \) is frequency of feature \( f \) that occurs in document \( d \),
- \( |D| \) is total number of documents in training set, and
- \( |DF_{(f)}| \) is the number of documents that have feature \( f \) occurring at least once.

2.1.2 Information Gain (IG): IG uses information gain algorithm to select features from training data. The algorithm will compute gain for each attribute by considering entropy [10]. The attribute with high value of gain shows that it is effective in classifying the training data. The following equations show entropy (2) and information gain (3) computation.

\[ Entropy(t) = -\sum_j p(j|t)\log_2 p(j|t) \]  

\[ IG = Entropy(p) - \left( \sum_{i=1}^{k} \frac{N_i}{n} \times Entropy(i) \right) \]  

2.1.3 Gain Ratio (GR): GR has been introduced to resolve the bias toward attributes with multi-values problem by adjusting information gain [10]. In text classification, GR is used to evaluate reliability of features by computing GR in each class with the value of information gain.

2.1.4 Chi Square (\( \chi^2 \)): \( \chi^2 \) is a statistical filtering approach used in feature selection to measure the degree of dependency between a features and a specific class [7]. The computation is as equation (4).

\[ CHI_{avg}(t) = \sum_{i=1}^{m} p(C_i) \times CHI(t, C_i) \]  

where
- \( CHI_{avg}(t) \) is chi square value of average of term \( t \),
- \( p(C_i) \) is probability values of class\( i \), and
- \( CHI(t, C_i) \) is chi square of term \( t \) in class\( i \).

2.1.5 ReliefF: ReliefF is a statistical-based method to evaluate feature in text classification. It uses weight of feature to consider relationship between a feature and specific class. This method is noise tolerant and also unaffected by feature interaction. Moreover, it can deal with multi-class problem, data redundancy, and missing values [11].

2.1.6 Support Vector Machine (SVM): SVM classifies data that are represented in a vector format. It builds linear equation to partition the dataset. There are two main processes of SVM. First, it starts by mapping non-linear data from input space to feature space. Then it uses kernel function to measure similarity in the feature vector [7]. The advantages of this method are the abilities to support large feature vector and to provide high accuracy model. The disadvantage of SVM is difficulty in selecting appropriate kernel function.

2.2 Related Work

There exist many works that comparatively study feature selection methods and classification algorithms in text classification. Some of them only had compared feature selection methods for a single classification algorithm [12] [13] [14]. Several researchers tried to compare multi feature selection methods and multi classification algorithms [15] [16] [17]. Feature selection methods are ranging from Information Gain, Gain Ratio, Chi Square,
Document frequency, TF-IDF, CFS, to many more. In text classification task, simple algorithms such Naïve Based was used widely in this task. Moreover, there are decision tree, Support Vector Machine (SVM), and k-nearest neighbor (KNN) that have been used in the task [18].

The specific task of categorizing places or point of interests (POIs) is divided in two groups of researchers. The first team [36] attempted to create a multilingual geographical gazetteer by using free data sources on the Internet. Data mining had important role to deal with large scale of data in this work, particularly text classification which are used to categorize geographical type by using place name as features and geographic object as classes. Experimental results displayed that place name categorization provided high precision but low recall.

The second team presented POI recommendation based on ontology [20]. A site categorization was implemented in this work. They used only sites’ name as features for learning with only Naïve Bayes algorithm to build classification model. The model was used to assign label to unknown sites. The experimental results showed that the model provides high accuracy of classification results, particularly the data in classes of restaurant, pharmacy, school and taxi.

As mentioned above, the experimental results in [19] and [20], it could be assumed that the categorizations were overfitting because both of them provided high accuracy, but low recall, of categorization. Thus, this work proposes the difference data source of place and considers only place’s description instead of their names to classify their classes. Furthermore, it also compares various feature selection approaches with different classification algorithms on this data.

3 Restaurant Classification Method

3.1 Overview Research Framework

In this work, the comparison of feature selections and classification algorithms for text classification to classify restaurant dataset is presented. The dataset is gathered from a famous travel website: LonelyPlanet.com. Only four classes of restaurants are considered. There are Thai, Chinese, Japanese and Italian restaurants. Overview (figure 1) of research framework constitutes the five main steps: (1) Data Collection, (2) Feature Extraction, (3) Feature Selection, (4) Classification, and (5) Models’ Performance Evaluation.

3.2 Data Collection

The collected restaurant data, both training and test data come from the LonelyPlanet website. Then they are stored in a database. The stored dataset includes description and class of restaurants. MYSQL is used as the database management software with PHP language to manage and query restaurant dataset in this work.

3.3 Feature Extraction

Main purpose of the second step is to transform the restaurant dataset in the database into a format suitable for data mining by extracting and reducing terms from text. Natural language processing (NLP) technique is used to extract and reduce terms from restaurants’ description. There are four processes of NLP used in this work:

a) Text Cleaning: All of restaurant descriptions are transformed into plain text. Then they are removed the HTML tags.

b) Word Parsing: The second process is splitting terms from text by determine space between each term as a separator.

c) Stop Words Removal: The third process is removing the segmented words or features that are not important such as article (“a”, “an”, “the”), pronoun (“he”, “she”, “it”), preposition (“in”, “on”, “under”) and conjunction (“and”, “but”, “also”). These words are called stop words.

d) Stemming: The fourth process is to convert terms into their original form without prefixes and suffixes. For example, the word “walk” can appear in different forms in many documents such as “walked”, “walking”, and “walks”. The stemming
technique can reduce occurrence frequency of these words in documents by transforming them into one word as “walk”. Porter Stemming algorithm is used to deal these terms in the restaurant dataset.

3.4 Feature Selection

There are large amount of extracted features from the feature extraction step. If classification models are created with all of them, they could give a low-accurate classification result on a test dataset, which is the effect known as overfitting. Therefore, feature reduction is used to be applied in the next step. There are two points which should be considered for feature selection: (1) which features should be selected to adequately represent the documents, and (2) how many features can give the most efficiency classification results. Hence, statistical filtering approaches are methods for feature selection in this work. Features with high statistical significance are selected for generating classification models.

The feature selections are divided to three approaches. The first approach considers TF-IDF measure to weigh features from each class and it selects only high significant features. The second approach combines TF-IDF with other statistical measures such as Information Gain, Gain Ratio, Chi Square, ReliefF, and SVM. The third approach, the method is the same as the second one but the quantity of features is smaller.

3.5 Classification

The selected features of training and test sets are transformed into a document matrix. Then only the document matrix of training set is used to create models with classification algorithms. The purpose of the fourth step is selecting the best algorithm for classifying the restaurant dataset. There are four algorithms, two come from Decision Trees (ID3 and C4.5) and other two come from Bayes' Theorem (Bayes Net and Naïve Bayes).

3.6 Models’ Performance Evaluation

After the classification models were built. They are used to evaluate their performance in the fifth step. The document matrix of test set from the previous step is taken to evaluate with the models. The performance evaluation is considered both number of features and accuracy of models.

4. Experimentation

4.1 Dataset

Dataset for the three experiments is the restaurant data gathered from LonelyPlanet.com website. Total number of data is 402 instances. It is separated into training and testing sets. The 205 instances of training set are restaurants in Asia and the 197 instances of a test set are restaurants in Europe.

4.2 Hardware and Software Specifications

All of experiments are tested with a personal computer with Microsoft Windows 7 Operating System. CPU speed is 2.30 GHz with 4 GB of Memory, 500 GB of storage and Gigabit Ethernet connection. The NLP techniques used in the feature extraction step, it is implemented with PHP version 5.2.6 and MySQL Database Version 5.0.51b. Weka version 3.6.5 for Windows OS is software for the data mining processes.

4.3 Experiment 1

The first experiment is to evaluate classification model with the first feature selection approach. The threshold values are set to select features from each class. The selected features are set of features which have TF-IDF values not less than the threshold. Thus, there are many ranges of features for testing, so the question is “which ranges can give the best accuracy”. The selected features from each class are combined into one set and merged duplicate words into one word.

4.4 Experiment 2

The second experiment uses the second feature selection approach. It differs from the first one in that the search of features is not in each class, but it finds features by considering them in all classes simultaneously. Therefore, the quantity of features in this method is more than in the first experiment. It totally has 278 features. The large number of features should be reduced. Thus, the other feature reduction approaches based on statistical measure are considered to reduce number of features. The amount of features of each feature reduction method is presented in Table 1.

4.5 Experiment 3

The third experiment uses the same method as the second one. Yet it reduces amount of features into equal quantity in each feature reduction methods. The number of features in each feature reduction methods is showed in Table 1.
Table 1: Number of features of each feature reduction method in the experiment 2 and 3

<table>
<thead>
<tr>
<th>Feature Reduction Methods</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original (TF-IDF)</td>
<td>278</td>
<td>278</td>
</tr>
<tr>
<td>Information Gain</td>
<td>200</td>
<td>40</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td>236</td>
<td>40</td>
</tr>
<tr>
<td>Chi Square</td>
<td>236</td>
<td>40</td>
</tr>
<tr>
<td>ReliefF</td>
<td>170</td>
<td>40</td>
</tr>
<tr>
<td>SVM</td>
<td>178</td>
<td>40</td>
</tr>
</tbody>
</table>

5. Results and discussions
In this section, the results of three experiments are described. There is also a discussion about the experimental results and suggestion on how to implement each feature selection approach with an appropriate classification algorithm.

5.1 Experimental Result 1
The accuracy of each classification algorithms with many ranges of features in the first experiment could be presented in figure 2.

![Figure 2: The first experimental result](image)

Figure 2 presented the four algorithms which could give the best accuracy in 41-51 feature ranges. C4.5 algorithm provided the highest accuracy at 89.34% with 41 features. ID3, Bayes Net and Naive Bayes Bayes algorithms gave the highest accuracy with 51 features at 85.79%, 87.82%, and 89.85%, respectively.

5.2 Experimental Result 2
The results in the second experiment (figure 3) showed that the decision trees algorithms could give significantly higher accuracy than the Bayes’ theorem algorithms. All feature reduction methods could increase accuracy, especially the accuracy of decision trees algorithms. C4.5 algorithm provided the highest accuracy at 85.79% with Information Gain and ReliefF feature selection methods. ID3 algorithm gave the best accuracy at 87.31% with Gain Ratio. Bayes Net and Naive Bayes algorithms gave the finest accuracy at 83.25% and 83.76% with ReliefF.

![Figure 3: The second experimental result](image)

5.3 Experimental Result 3
The third experimental result in figure 4 showed that the accuracy from two algorithms of Bayes’ theorem were higher than those of decision trees in every feature reduction methods. C4.5 algorithm gave the best accuracy at 85.79% with Information Gain, ReliefF and Chi Square selection methods. ID3 algorithm provided the highest accuracy at 86.80% with Gain Ratio. Bayes Net gave the highest accuracy at 89.34% with Gain Ratio. Naive Bayes gave the highest accuracy at 89.85% with ReliefF.

![Figure 4: The third experimental result](image)

5.4 Discussion
The study compared the text classification performance for classifying restaurant dataset. The classification considered restaurants’ description to assign their classes by using NLP and data mining technique. The performance of text classification depends on number of features and classification algorithms. The three feature selection approaches and four classification algorithms were applied in our study.

The feature selection of the first approach took longer time than other approaches because it selected the features from each class. The result of...
the first experiment indicated that the most suitable quantity of feature was around 40-50 features and the highest accuracy got from C4.5 and Naïve Bayes.

The feature selection methods in the second and third approaches took shorter time than the first one, because they selected the features from all class in one time. Since, the poor feature selection of the second approach, dominant features in some classes were neglected resulting in a poor predictive performance of classification model. In case of the reduced features are still at high quantities, the algorithms that could provide good accuracy are decision trees. The third experiment result indicated that when the features were reduced to small quantity, algorithms from Bayes’ theorem could give higher accuracy than decision trees.

6 Conclusion & Future Work

The purpose of this paper is to provide criteria to select the most suitable features and the appropriate algorithm for restaurant data classification. There are four classes of restaurants that are collected from the website “LonelyPlanet.com”. Restaurant description is the attribute for classifying. Natural language processing has been applied in the feature extraction step. The feature selection step uses statistical approaches such as TF-IDF, Information Gain, Gain Ratio, Chi Square, ReliefF and Support Vector Machine, to select significant features. There are four classification algorithms applied in this work: C4.5, ID3, Bayes Net, and Naïve Bayes.

The experimental results indicated that with large number of features, C4.5 and ID3 models can provide higher accuracy than Bayes Net and Naïve Bayes models. But with the small amount of features, algorithms based on Bayes’ theorem can give higher accuracy than algorithms based on decision tree. In the future, we plan to implement an application to classify restaurant data from multi data sources.

References: