Text Categorization using
a New Text Association Rule-Based Classifier

Supaporn Buddeewong and Worapoj Kreesuradej
Faculty of Information Technology
King mongkut’s Institute of Technology Ladkrabang,
Bangkok, 10520 Thailand.

Abstract. This paper proposes a new association rule-based Classifier algorithm to improve the prediction accuracy of Association Rule-based Classifier By Categories (ARC-BC) algorithm. Unlike the previous algorithms, the proposed association rule generation algorithm constructs two types of frequent itemsets. The first frequent itemsets, i.e. $L_k$, contain all term that have no an overlap with other categories. The second frequent itemsets, i.e. $OL_k$, contain all features that have an overlap with other categories. In addition, this paper also proposes a new join operation for the second frequent itemsets. The experimental results are shown a good performance of the proposed classifier.

Keywords. Text Categorization; Text mining; Classification; Association rule.

1. Introduction

Text categorization task is defined as assigning category labels to new documents based on their contents. Text categorization research has a long history, starting in early 1960s. There are several text categorization approaches have been proposed such as neural networks, genetic algorithms and probabilistic models, support vector machine.

Recently, Association Rule-based Classifier By Categories (ARC-BC) algorithm [4, 5] that is based on association rule mining approach have been proposed. These classifiers have been proven to powerful. In addition, these classifiers produce clear and understandable results. However, an algorithm can not work well for the single-class document that has some terms of document mutually associated with other class. As a result, an algorithms may incorrectly classify those single-class documents. Therefore, this paper proposes a new association rule-based classifier algorithm to improve the prediction accuracy of ARC-BC algorithm.

This paper is organized as following. The second section gives the overview of a text categorization with association rule. In the third section, we introduce our new text categorization approach. Experimental results are described in Section 4. We summarize our research and future work in the fifth Section.

2. Text categorization with association rule

The construct process of an associative classifier is shown in figure 1.

Here, the training set is a document collection. A document $D_i$ of the collection is assigned to a set of categories $C = \{c_1, c_2, c_3, ..., c_m\}$. At preprocessing phase, The set of term $T = \{t_1, t_2, t_3, ..., t_n\}$ of document $D_i$ is retained after term pruning and stemming. Then, a document $D_i$ is model as the following:

$D_i = \{c_1, c_2, .., c_m, t_1, t_2, ..., t_n\}$.

Figure 1. Association classification method [5]

The next step is the generation of association rule. At the associative rule mining, a set of rules that associate the terms of a document and its categories is extracted from the training data by using an apriori-based algorithm. However, the association rules are constrained in that the antecedence has to be a conjunction of term from $T$, while the consequence of the rule has member of $C$.

After generating the set of rule, an important step is building and validating an associative text classifier. Recently, Association Rule-based Classifier By Categories (ARC-BC)
algorithm is proposed to build an associative text classifier. Method of ARC-BC considers each set of documents belonging to one category as a separate text collection to generate association rules. If a document belongs to more than one category, this document will be present in each set associated with the categories that the document falls into. However, both algorithms may fail to classify a single-class document that has some terms of document mutually associated with other class. Therefore, in this paper, we propose a new text rule-based classifier to deal with such problem. The new text association rule-based classifier is presented in the next section.

3. Categorization using a new association rule-based classifier

A new algorithm for a association rule-based classifier is introduced. Our algorithm is proposed to deal with misclassifying problem of a single-class document that has some terms of document mutually associated with other class. Here, a new algorithm for association rule generation and a new categorization algorithm based on the new set of rules is proposed.

3.1. Association rule generation

Like typical apriori-based algorithms, our associative rules are generated from frequent itemsets. Therefore, frequent itemsets are constructed in order to get the set of associative rules. However, unlike the previous algorithms, our method constructs two types of frequent itemsets. The first frequent itemsets, i.e. \( L_1 \), contain all term that have no overlap with other categories. The second frequent itemsets, i.e. \( OL_k \), contain all features that have an overlap with other categories.

Basically, any frequent itemsets has to include a category label starting from 2-itemses to k-itemses. For the first frequent itemsets, \( L_1 \), is constructed using apriori join. As an example, \( L_4 \) is constructed by using apriori join between \( L_4 \) and \( L_{k-1} \). In this paper, a new join operation for the second frequent itemsets, \( OL_k \), is proposed.

For the new join operation, any two items in an itemset can be joined if they have the same category. The examples of both apriori join shown in figure 2 and the new join shown in figure 3. For figure 2, figure 3 and figure 4. We use the model of document as

\[
D_j = \{ c_1, c_2, c_3, a, b, c, \ldots, z \}.
\]

Document \( D_j \) consists of category label, \( c_1, c_2, c_3 \). Term of document are represent by character, a-z.

In addition, the algorithm, called ARTC, for generating association rules are shown in figure 4.

![Figure 2. Apriori join](image)

![Figure 3. ARTC join](image)

![Figure 3. Sample of T, L2, OL2 and ML](image)

![Figure 4. Algorithm for generating association rule](image)
3.2. Prediction of classes association with new document

After obtaining a set of associative rules, an associative classifier makes use of the set of associative rules in the prediction of classes for new documents. The algorithm for classifying a new document is shown in figure 5. Basically, the algorithm takes feature terms from each new document to compare with each rule in the set of rules. If all feature terms are matched with the rule, that new document is classified by that rule.

\[
\begin{align*}
(1) & \text{for each new document} \\
(2) & R \leftarrow \emptyset \\
(3) & \text{for each Rule} \\
(4) & \text{if feature_new_doc } \subseteq \text{ Rule} \\
(5) & R = \text{ rule} % rule } \in \text{ Rule} \\
(6) & \text{end} \\
(7) & \text{group R by category: R_1, R_2, ..., R_n} \\
(8) & \text{for each R} \\
(9) & \text{sum the confidences of rules and divide by the number of rules in R} \\
(10) & \text{sum the support of rules and divide by the number of rules in R} \\
(11) & \text{end} \\
(12) & \text{if 1 group has highest number of rule} \\
(13) & \text{put the new document in the class that has highest number of rule in group.} \\
(14) & \text{if more 1 group has highest number of rule} \\
(15) & \text{put the new document in the class that has highest confidence sum.} \\
(16) & \text{if more 1 group has highest confidence sum} \\
(17) & \text{put the new document in the class that has highest support sum.} \\
(18) & \text{end} \\
(29) & \text{next new document}
\end{align*}
\]

Figure 5. Algorithm for classifying a new document.

4. Experimental results

In this section, the experimental results of document classification using text rule-base classifier are reported.

For our experiment, we use data generated dataset and Computer Science Technical Report collection dataset (CSTR) dataset for test our algorithm. To evaluate the proposed algorithm, we use classification accuracy rate of classification.

Data generated dataset is the text data that we generated for basic testing our algorithm. The dataset divided into 3 category: Category1, Category2 and Category3. Venn diagram of feature term is shown in figure 6. We generated 3 dataset, each dataset contained 1000 documents. For each dataset, the training data consists of 800 documents and testing data consists of 200 documents. Some of data generated documents are shown in table 1.

The table 2 showed the accuracy rate of classification by ARC-BC and ARTC algorithm. The accuracy rate from ARTC algorithm is more than that from ARC-BC algorithm.

CSTR data set is the abstracts of URCS technical reports published in the department of computer science at the University of Rochester between 1987 and 2005. It has been use in [1, 3, 4, 5] for text categorization and clustering. The dataset contained 472 abstracts, which were divided into four research areas: AI, Robotics and Vision, Systems and Theory. The set of keywords of a documents is used on the set of term, \( T = \{t_1, t_2, ..., t_n\} \), of document.

CSTR document is shown in figure 7. It’s an abstract from 472 documents. Each document consists of title keywords and abstract.

![Venn diagram of features term](image1)

![Figure 6. Venn diagram of features term](image2)

![Figure 7. Venn diagram of features term](image3)
Keywords: temporal invariance; predictive coding; unsupervised learning.

Predictive coding and temporal invariance are two major unsupervised learning principles which have been used to explain the behavior of parts of the brain (most notably the striate cortex). Although both have been around for a number of years, no formal relationship between them has been established. We prove that temporal invariance is a form of predictive coding. To do this, we begin with the goal of predictive coding, make a set of assumptions about the class of problem we are dealing with, and derive temporal invariance from the predictive coding goal and our added assumptions.

CSTR experimental results are shown in table 3.

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARC-BC</td>
<td>ARTC</td>
</tr>
<tr>
<td>AI</td>
<td>65.22</td>
</tr>
<tr>
<td>Systems</td>
<td>87.88</td>
</tr>
<tr>
<td>Robotics and Vision</td>
<td>70.37</td>
</tr>
<tr>
<td>Theory</td>
<td>89.74</td>
</tr>
<tr>
<td>All</td>
<td>80.33</td>
</tr>
</tbody>
</table>

The table 3 shows the accuracy rate of classification by categories that are given by ARC-BC algorithm and ARTC algorithm. In addition, the last line from table 3 also shows the total accuracy rate of the CSTR dataset. The accuracy rate from ARTC algorithm is more than that from ARC-BC algorithm.

5. Conclusion and future work

In this paper, we introduce the new algorithm for text categorization, Association Rule-Based Text Classifier algorithm (ARTC). According to the experimental results, ARTC algorithm shows good performance in classifying the data. In the future, we will further conduct experimentations on k-datasets and Reuters-top10 collection datasets to evaluate the performance of the algorithm.

6. References