Adapting Associative Classification to Text Categorization

Baoli Li[1], Neha Sugandh[1], Ernest V. Garcia[2], Ashwin Ram[1]

[1] College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
+1(404)385 1186
{baoli, nsugandh, ashwin}@cc.gatech.edu

[2] Department of Radiology
School of Medicine, Emory University
Atlanta, GA 30322, USA
+1(404)353 0143
Ernest.Garcia@emoryhealthcare.org

ABSTRACT
Associative classification, which originates from numerical data mining, has been applied to deal with text data recently. Text data is firstly digitalized to database of transactions, and then training and prediction is actually conducted on the derived numerical dataset. This intuitive strategy has demonstrated quite good performance. However, it doesn't take into consideration the inherent characteristics of text data as much as possible, although it has to deal with some specific problems of text data such as lemmatizing and stemming during digitalization. In this paper, we propose a bottom-up strategy to adapt associative classification to text categorization, in which we take into account structure information of text. Experiments on Reuters-21578 dataset show that the proposed strategy can make use of text structure information and achieve better performance.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information Filtering*; I.5.4 [**Pattern Recognition**]: Applications – *Text processing*.

General Terms
Algorithms.

Keywords
Text Categorization, Associative Classification.

1. INTRODUCTION
Associative classification, which integrates association rule mining and classification, is originally proposed in data mining community. Generally speaking, this kind of approach firstly generates a complete set of association rules, and then chooses a small set of high quality rules for prediction.

In the past few years, several effective and efficient associative classifiers have been developed, such as CBA [1], CMAR [2], CPAR [3], HARMONY [4], and so on. Besides numerical transactional datasets, they have also been applied to process text data [4, 5]. However, when dealing with text categorization problem, they usually follow an intuitive strategy: transform text data into transactional dataset and then apply associative classifier on the derived dataset as usual. Although an associative document classifier has to deal with some specific problems of text data such as lemmatizing and stemming during digitalization stage, this widely used strategy does not pay much attention to the inherent characteristics of text data.

For example, a document is usually organized into a hierarchical structure of content units with different granularities, from individual word to the whole document. It is well known that much semantic information is expressed by text structure, rather than by individual words. Unfortunately, the current strategy ignores most structure information in text. This can be observed from the format of the associative classification rules generated by the current strategy. For example, the following rule is derived from Reuters-21578 dataset: "offer, share, tender -> acq". We can imagine that the antecedent of this rule may have some substructures. Two of these three words, e.g. offer and share, may occur often in the same sentences. Thus, a more accurate classification rule with substructure like "(offer, share), tender -> acq" could be obtained. This kind of rules with substructures, which can be obtained based on text structure, is expected to be helpful to improve the performance of an associative classifier for text categorization.

In this research, we try to design such a new strategy that could make use of text structure information to make associative classification more applicable to text categorization. We propose a bottom-up model to derive associative classification rules with substructures for building a classifier with much better discriminative power.

The paper is organized as follows: section 2 details our proposed bottom-up strategy. Section 3 gives experiments and discussions on Reuters-21578 dataset. We conclude the paper with future work in section 4.

2. OUR PROPOSED STRATEGY
Associative classification rule $\{t_1, t_2, \ldots, t_k\}$->$c$ is a special type of association rule, where the rule's consequent c is a category label, and t in the antecedent stands for a term. Its formal description is given as follows:

ACR ::= A -> C

A ::= A, T | T
T ::= t_1 | t_2 | … | t_n
C ::= c_1 | c_2 | … | c_m

Where T is the set of terms, and C is the set of category labels. Similarly, associative classification rule with substructures can be formally described as follows:

ACR_S ::= A->C
A ::= A, A | (A) | T
T ::= t_1 | t_2 | … | t_n
C ::= c_1 | c_2 | … | c_m

To derive such a kind of structured classification rules, we propose a bottom-up strategy. It is a natural solution as a document is created recursively from lower and smaller content units to higher and larger ones. The minimal content unit is *word*, where the maximal unit is *document*. Other content units from lower to higher include: *phrase*, *clause*, *sentence*, *paragraph*, *section*, and so on.

We can start from any content unit level to build a structured associative classifier. At the lowest level, a term corresponds to a word, and a base associative classifier is firstly applied to obtain a classification model, which consists of a set of unstructured classification rules. At a higher level, a term can be not only a word, but also a compound term that is directly derived from the antecedent of an unstructured or structured classification rule obtained in the immediately previous level. This process will run recursively until we reach the highest *document* level.

Actually, in this bottom-up strategy, the lower level classification models are used to generate compound terms for the higher levels. These compound terms are thought to convey more specific and accurate semantic information than individual terms. In this sense, the lower level models work as a cascaded feature generation and selection module for the higher level models. However, the process of feature selection is done by a set of cascaded associative classifiers rather than by some statistical metrics like information gain.

The proposed bottom-up strategy is flexible and scalable, because:

1. We can use different existing associative classifiers as base classifier. At different levels, we can use different base classifiers.
2. In theory, we can start from any content unit level: word, phrase, clause, sentence, paragraph, or section. However, as the length of a compound term is hopefully greater then one, we'd better start from levels other than *word* in practice. If we begin with *document* level, our classification strategy will behave as the chosen base associative classifier. We will obtain classification rules without substructures.
3. We don't need to go through all levels from the chosen start level to the final *document* level. We can choose the required levels from the candidate set according to the dataset to be processed. For example, we may consider only two levels: *sentence* and *document*, and ignore other intermediate levels. However, *document* level is a must as the final one.

Compared to the traditional associative classification rules, the structured rules derived by this strategy are relatively easier to understand, interpret, and revise if needed.

3. EXPERIMENTS AND DISCUSSIONS

We tested our proposed strategy on the ModApte split version of Reuters-21578 dataset (http://kdd.ics.uci.edu/databases/reuters2-1578/reuters21578.html). Like many other studies, the top 10 categories containing the most number of documents were used in our experiments, which results in a training set with 7,193 documents and a test set with 2,787 documents. These 10 categories and their distribution in the training set are listed in the first column of table 1.

The instance-centric associative classifier HARMONY [4] is chosen as base classifier for all levels in our experiments, as it demonstrates quite good accuracy and efficiency on both numerical transactional databases and text datasets. As *sentence* is the most easily recognized unit in text, we choose it and *document* to build a two-level model in our experiments.

At the preprocessing stage, we remove stop words and stem surface words. To split sentences, we use a simple strategy based on punctuation like "." and "?". For the parts without such punctuation, we regard each line as a sentence.

The harmonic mean of Precision and Recall, i.e. F-1 measure, is used for measuring performance. To verify the effectiveness of our proposed strategy, we experiment with the following methods:

HMY: the original HARMONY algorithm.
ADPT_HMY: our proposed strategy with the original HARMONY algorithm as base classifiers on both sentence and document levels.
***k*NN**: *k*-nearest neighbor algorithm. The parameter *k* was set to be 30.
SVM: probabilistic Support Vector Machine algorithm with linear kernel. We use the LIBSVM[1] package in our experiments.

For HMY and ADPT_HMY, we use the same absolute support threshold 80, and an entropy based discretization method [6] is used to discretize words on their frequencies.

For *k*NN and SVM, we use the following equation to calculate the weight W_{t_i,d_j} of term t_i in a document d_j.

$$W_{t_i,d_j} = (\log(TF_{t_i,d_j}) + 1) \times \log(\frac{N}{DF_{t_i}}) \qquad (1)$$

Where *TF* is the count of t_i's occurrence in document d_j, *DF* is the number of documents in which the term t_i occurs, and *N* is the total number of documents in the collection.

Table 1 shows the results of four methods on the top 10 populated categories. Our adapted strategy ADPT_HMY performs quite well. The Micro-Averaged and Macro-Averaged F-1 measures are about 1.86% and 3.27% higher than those of the original HARMONY algorithm. Its micro-averaged F-1 measure is comparable to that of SVM, which is recognized to be the winner on many datasets. The relatively lower macro-averaged values for these four methods are due to the unbalanced category distribution in the dataset. Compared to *k*NN and SVM, HMY and ADPT_HMY achieve much better macro-averaged scores and

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

perform quite well on small categories, as can be attributed to the instance-centric rule-generation strategy of HARMONY. On the contrary, SVM demonstrates excellent generalization capacity on large categories. Among the 10 categories, SVM achieves the best performance for five large categories, where ADPT_HMY and HMY perform best for three and two small categories respectively. For documents of category "Corn", SVM incorrectly assigns all of them other category labels, which makes SVM achieve the poorest performance on this rare category.

Table 1. F-1 measures for the ten most populated categories of Reuters-21578 (ModApte split version) (%).

Algorithm / Category	HMY	ADPT_ HMY	*k*NN (*k*=30)	SVM (linear)
Acq (22.94%)	90.20	93.25	89.32	**97.08**
Corn (2.52%)	34.29	**37.14**	9.37	0.00
Crude (5.41%)	83.43	84.21	79.22	**84.32**
Earn (40%)	93.97	96.35	94.33	**98.90**
Grain (6.02%)	**90.84**	85.61	52.00	60.33
Interest (4.82%)	62.61	**72.41**	61.40	71.49
Money-fx (7.48%)	77.28	76.74	75.62	**79.27**
Ship (2.74%)	**75.90**	73.07	62.00	65.52
Trade (5.13%)	78.26	76.76	75.70	**87.24**
Wheat (2.95%)	45.54	**69.42**	39.26	30.77
Macro-Average	73.23	**76.50**	63.82	67.49
Micro-Average	86.90	**88.76**	82.60	88.55

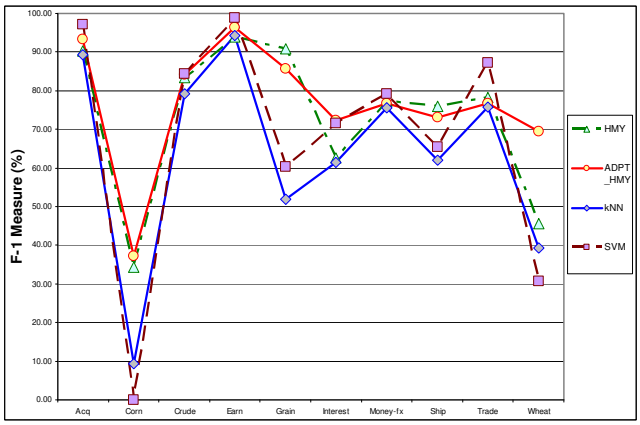

Figure 1. F-1 measures of the top 10 populated categories on Reuters-21578 (ModApte split version).

Figure 1 visualizes the experimental results. It clearly shows that ADPT_HMY is the most stable algorithm on the top 10 populated categories. The standard deviations of the four methods on these 10 categories are 0.1994 (HMY), 0.1648 (ADPT_HMY), 0.2542 (kNN), and 0.31 (SVM), respectively. The hardest category for these four methods is "Corn". The two associative classification methods (HMY and ADPT_HMY) obtain relatively better performance on this category, which demonstrates their advantages on confusing and small categories.

Figure 2 provides 5 example rules derived by HMY and ADPT_HMY. (H1) and (H2) are learned by HMY method, while others are generated by ADPT_HMY. (A3) and (A4) are the structured version of (H1). Similarly, (A5) corresponds to (H2), but with substructures. The confidence values of the structured rules are higher than those of their corresponding unstructured ones. The better performance of ADPT_HMY indicates that the associative classification rules with substructures have better discriminative power than those unstructured ones.

(H1) dividend, record, declare -> earn (conf=0.982)

(H2) currenc, exchange, rate -> money-fx (conf=0.72)

(A3) (dividend, record),(dividend, declare) -> earn (conf=1.0)

(A4) (dividend, record), declare -> earn (conf=0.989)

(A5) (exchange,rate),(currenc,rate) -> money-fx (conf=0.787)

Figure 2. Examples of derived associative classification rules.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a bottom-up strategy to make associative classification more applicable to text categorization. Compared to the previous intuitive one, this strategy can make use of text structure information and thus derived classification rules have better discriminative power. In the future, we will experiment this strategy with more datasets and other base associative classifiers. We are also planning to apply this strategy with other machine learning algorithms for text categorization.

5. REFERENCES

[1] Liu B., Hsu W., and Ma Y. Integrating classification and association rule mining. In *Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'*98), New York, NY, August 1998.

[2] Li W., Han J., and Pei J. CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In *Proceedings of IEEE International Conference on Data Mining* (*ICDM '01*), 2001.

[3] Yin X. and Han J. CPAR: Classification based on predictive association rules. In *Proceedings of SIAM International Conference on Data Mining* (*SDM*'03), San Francisco, CA, May 2003.

[4] Wang J. and Karypis G. Harmony: Efficiently mining the best rules for classification. In *Proceedings of SIAM international conference on Data Mining Proceedings* (*SDM'05*), 2005.

[5] Antonie, M.-L. and Zaiane, O. R. Text Document Categorization by Term Association. In *Proceedings of IEEE International Conference on Data Mining* (*ICDM '02*), 2002.

[6] Fayyad, U. M. and Irani, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of International Joint Conference on Artificial Intelligence* (*IJCAI-1993*), 1993.