International Journal of Intelligent Information and Management Science

Volume 4, Issue 1, February 2015

http://www.hknccp.org

Publisher: HongKong New Century Cultural Publishing House

Address: Unit E79, 3/F., Wing Tat Commercial Building, 97 Bonham Strand East, Sheung Wan, HK.

Contents

H K . N C C P

International Journal of Intelligent Information and Management Science
ISSN: 2307-0692 *Volume 4, Issue 1, February 2015*

Short Text Categorization Based on Instance Selection Learning

Yan ZHAN, Hao CHEN, Guochun ZHANG
College of Mathematics and Information Science, Hebei University, Baoding, 071002, CHINA

Abstract: Text Categorization is an important component in many information organization and information management tasks. In short text classification problem, which is as a branch of text classification, there will be too many instances which need much computation time and memory requirement. This paper proposes an instance selection learning method which can reduce the instance numbers in K-NN classification algorithm. The experiments also compared the learning algorithm with existing reducing samples algorithms such as Condensed Nearest Neighbor, Selective Nearest Neighbor, Reduced Nearest Neighbor Rule, Edited Nearest Neighbor Rule in Short Text Categorization.

Keywords: Short Text Categorization; Instance Selection Learning; K-NN Classification

1. Introduction

The automated categorization of texts into topical categories has a long history, dating back at least to 1960. Until the late ’80s, the dominant approach to the problem involved knowledge-engineering automatic categorizers, i.e. manually building a set of rules encoding expert knowledge on how to classify documents. In the ’90s, with the booming production and availability of on-line documents, automated text categorizations has witnessed an increased and renewed interest [1].

Text Categorization (TC) is an important component in many information organization and information management tasks.

Short text classification means the classification of few content texts (usually less than 100 words). Nationally and internationally text categorization [1,2] has many years of research, but in this essay classification field work less [3,4].

This short text classification problem as text classification a branch, in addition to the same with traditional text classification to a certain degree, still need to face some special problems to be solved, because text length short, features sparse, hard to measure the essay, similarity between simply from common text classification task of transplantation is often and can't get a good result [5].

Most algorithms to train artificial neural networks or machine learning methods use all vectors from the training dataset. However, there are several reasons to reduce the original training set to smaller one. The first of them is to reduce the noise in original dataset because some learning algorithms may be noise-fragile (for example, plain linear discrimination methods [6]). The second reason to shrink the training set is to reduce the amount of computation, especially for instance-based learning (or lazy-learning) algorithms [7] such as the K-nearest neighbors [8], or for huge training sets. The third and relatively new reason to use vector selection appeared together with new prototype selection algorithms. These algorithms shrink training sets sometimes even below 1% of original size keeping the accuracy for unseen vectors high. As the results of shrinking good prototype vectors are selected.

2. K-nearest Neighbor Algorithm

K-nearest neighbor is a classification method based on statistic theory. It is a method frequently used in data mining classification algorithm. This algorithm assumes all instances correspond to points in the n-dimensional space R^n. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance. More precisely, let an arbitrary instance x be described by the feature vector <a_1(x), a_2(x)…a_n(x)>, Where a_r(x) denotes the value of the rth attribute of instance x. Then the distance between two instance x_i and x_j is defined to be $d(x_i, x_j)$, where

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{1}$$

Learning in this algorithm consists of simply storing the presented training data. Each time a new query case is encountered, its relationship to previously stored cases is examined in order to assign a target function value for the new case. The larger the case base, the greater the problem space covered, however, it would also reduce the system’s performance if the number of cases exceeds an unacceptably high threshold. In general, in one case base, the denser the cases, the more redundant cases are contained. It is necessary to refine the case-base by deleting the redundant cases to improve the performance of NN algorithm [9].

3. Some Samples Reduction Algorithms

(1) Condensed Nearest Neighbor (CNN).

Hart made one of the first attempts to reduce the size of the training set with his Condensed Nearest Neighbor Rule (CNN) [10]. His algorithm finds a subset S of the training set T such that every member of T is closer to a member of S of the same class than to a member of S of a different class.

(2) Selective Nearest Neighbor (SNN).

Ritter et al. extended the condensed NN method in their Selective Nearest Neighbor Rule (SNN) such that every member of T must be closer to a member of S of the same class than to any member of T (instead of S) of a different Class [11].

(3) Reduced Nearest Neighbor Rule (RNN).

Gates introduced the Reduced Nearest Neighbor Rule [12]. The RNN algorithm starts with S=T and removes each instance from S if such a removal does not cause any other instances in T to be misclassified by the instances10 remaining in S.

(4) Edited Nearest Neighbor Rule (ENN).

Wilson developed the Edited Nearest Neighbor (ENN) algorithm in which S starts out the same as T, and then each instance in S is removed if it does not agree with the majority of its k nearest neighbors (with k=3, typically) [13]. This edits out noisy instances as well as close border cases, leaving smoother decision boundaries.

(5)Generalization Capability (GC) Algorithm.

We apply instance selection learning, which will reduce the case number into K-NN algorithm so that we can improve indexing efficiency in searching near neighbors. Firstly we introduce a concept called Generalization Capability of a case. Based on this concept, an approach to delete the redundant cases is presented. The Generalization Capability of cases indicates the problem-solving ability of cases. According to the proposed approach, the cases with better Generalization Capability are maintained as the representative cases in the case-base while those redundant cases found in their coverage are removed. The experiments show that the proposed method can greatly remove the redundant cases or less useful cases as well as preserve a satisfying accuracy of solutions when it is used for classification tasks [14].

An intuitive and powerful heuristic algorithm is designed below.

Step 1. The set R is initialized to be empty, $S = X \cup Y$.

Step 2. For each case x in X, determine Coverage (x) by the following (2).

Step 3. Find case x* such that $|Coverage(x^*)|$ = $Max_{x\in S}|Coverage(x)|$. If there exists more than one case such that the maximum is reached, select a case x** from them such that r_x is maximal. If here exists more than one case such that the maximum is reached, select one randomly.

Step 4. Put $R = R \cup \{x^*\}$ and $S = S - Coverage(x^*)$, if $|S| = 0$ then stop else go to Step 2.

Consequently, the set R is approximately regarded as the set of selected representative cases. He entire positive set X. In fact, the selected cases can exclude the entire negative set Y yet.

The coverage of x, Coverage (x), is defined as

$$Coverage(x) = \{e \mid d(x,e) < r_x\} \qquad (2)$$

We know that a positive example covers a negative example is impossible. In other words, the coverage of a positive case is a subset of X. The coverage of a case p represents the generalization capability of this case. The bigger is the number of cases in its coverage, the more representative is the selected case x. From (2), we can determine the coverage for each positive case p. The current objective is to select a set of positive cases such that the selected cases can cover the entire positive set X. In fact, the selected cases can exclude the entire negative set Y yet [15].

4. Experiments in Short Text Categorization

Many of the reduction techniques surveyed in Section 3 and all of the techniques proposed in Section 4 were implemented and tested on 600 papers from People's Daily, which belong to computer programming (CP), International News (IN), profile interview (PI), military (MI), women (WO) and sports (SP) news etc. And preprocess for the testing samples, keep only nouns, verbs and adjectives.

The basic K nearest neighbor (KNN) algorithm that retains 100% of the training set is also included for comparison. All of the algorithms use K = 3. (Experiments were also done using a more traditional Euclidean distance metric with overlap metric for nominal attributes.)

In this section, we apply GC, which will reduce the case number into K-NN algorithm so that we can improve indexing efficiency in searching near neighbors. According to the proposed approach, the cases with better Generalization Capability are maintained as the representative cases in the case-base while those redundant cases found in their coverage are removed. The experiments show that the proposed method can greatly remove the redundant cases or less useful cases as well as preserve a satisfying accuracy of solutions when it is used for classification task.

Ten-fold cross-validation was used for each experiment. We partition the 600 papers into five dataset. The average accuracy is reported for each reduction algorithm on each dataset in Fig. 1. The storage percentage is reported in Fig. 2.

To evaluate the effectiveness of a text categorization system, we use the standard recall, precision and F1 measure.

Recall(R) is defined to be the ration of the number of the correctly assigned documents to the number of positive samples. Precision (P) is the ratio of the number of the correct documents in the positively assigned documents. The F1 measure combines recall and precision in the following way:

$$F1=\frac{2*R*P}{R+P} \qquad (3)$$

In our experiments each data set was first partitioned into two sets of mutually exclusive randomly selected examples to form 90% of the databases as training data and 10% of the databases as testing data.

Experiment 1 (Exp.1) uses the traditional mutual information to make feature selection and k-NN classification. Experiment 2 (Exp.2) uses instance selection learning method based on Generalization Capability (GC) algorithm. From Fig. 3- Fig. 5 we can see experiment 2 improved the Precision, Recall and F1 value. Therefore, considering the instance selection learning, it can get better classification performance compared to the conventional the results of the experiment.

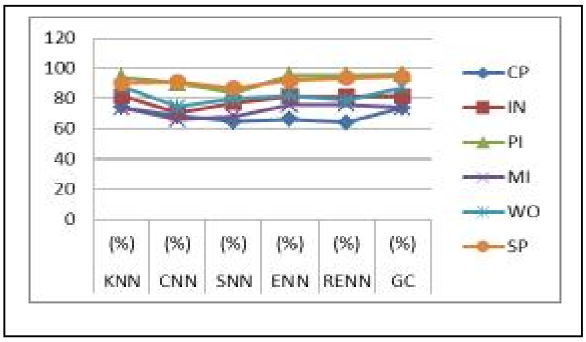

Figure 1. The accuracy comparison between KNN, CNN, SNN, ENN, RENN and GC

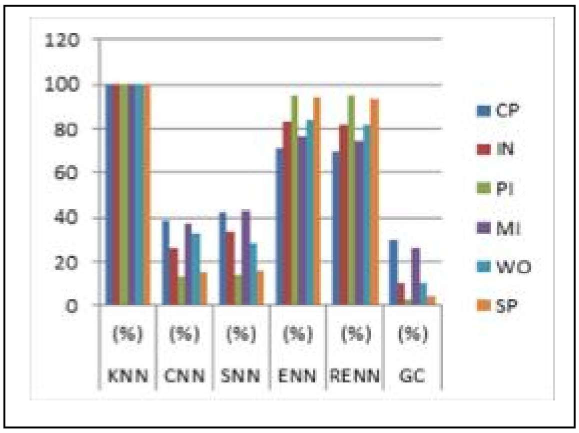

Figure 2. The Instance numbers between KNN, CNN, SNN, ENN, RENN and GC

As expected, ENN, RENN and K-NN all retained over 75% of the instances, due to their retention of internal (non-border) instances. They all had fairly good accuracy, largely because they still had access to most of the original instances. The GC method achieved better reduction and higher accuracy than RENN, which in turn had higher reduction than ENN. ENN's accuracy was significantly higher, but this is mostly due to retaining most of the instances. The GC, ENN and RENN methods also achieved higher accuracy than KNN, since they were designed specifically for noise filtering. They also required about 10% less storage than in the noise-free case, probably because they were throwing most of the noisy instances (as well as a few good instances that were made to appear noisy due to the added noise).

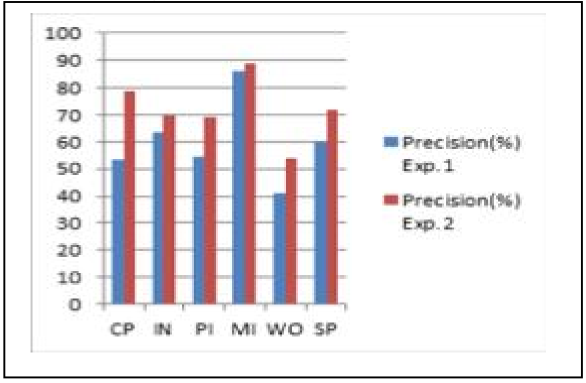

Figure 3. The Precision between Exp.1 and Exp. 2

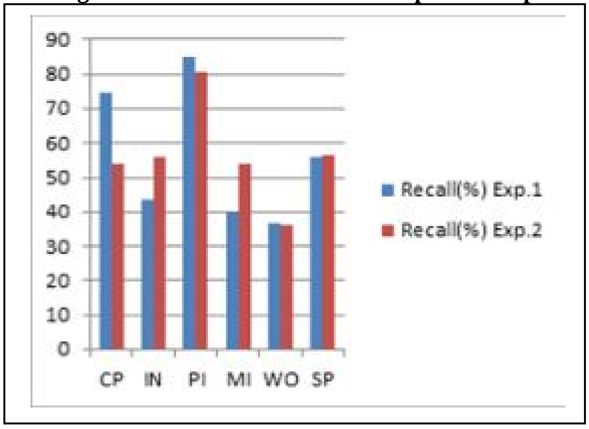

Figure 4. The Recall between Exp.1 and Exp. 2

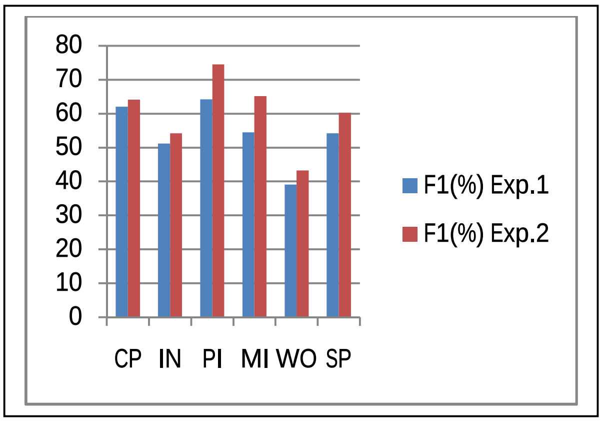

Figure 5. The F1 between Exp.1 and Exp. 2

5. Conclusion

In this paper short text categorization based on instance selection learning is described. It also compared the GC algorithm with existing reducing samples algorithms such as Condensed Nearest Neighbor, Selective Nearest Neighbor, Reduced Nearest Neighbor Rule, Edited Nearest Neighbor Rule in Text Categorization. The experiments verify that GC had significantly higher accuracy and less instance numbers than other algorithms. Therefore, considering the instance selection learning, it can get better classification performance compared to the conventional the results of the experiment. Since K-NN algorithm is used extensively to a variety of areas, we can improve classification performance further and makes its widespread application in short text categorization more valuable by optimizing this algorithm.

6. Acknowledgments

This work is supported by NSF of Hebei Province (No. F2013201064, F2014201100) and the Soft Science Research Program of Hebei Province (Grant no. 14450318D)

References

[1] Fabrizio Sebastiani, A Tutorial on Automated Text Categorisation, Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence (1999), pp. 7-35, Buenos Aires, AR.

[2] SebastianiI F.Machine Learning in Automated Text Categorization Consiglio Nazionale delle Ricerche.Italy.ACM Computing Surveys,2002,34(1), pp. 1-47.

[3] Fan Xing-hua, Sun Mao-song. A High-performance Text Categorization of Two Classes. Chinese Journal of Computer, 2006,29 (1), pp. 124-131.

[4] Zelikovitz S,Transductive M F.Learning for Short-Text Classification Problem using Latent Semantic Indexing International.Journal of Pattern Recognition and Artificial Intelligence,2005,19 (2), pp. 143-163.

[5] Yan Zhan, Hao Chen. Feature Extended Short Text Categorization Based on Theme Ontology, 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery. pp. 719-722.

[6] Duda, R.O., Hart, Pattern Classification and Scene Analysis. 2edn. Wiley.1997.

[7] Aha, D.W., Kibler, D.,Aha,. Machine Learning 6, pp. 37-66 (1991).

[8] Cover, T.M., Hart, P.E., Nearest neighbor pattern classification, Institute of Electrical and Electronics Engineers Transactions on Information Theory 13, pp. 21-27 (1967) .

[9] H.B. Mitchell, P.A. Schaefer, A "soft" K-Nearest Neighbor Voting Scheme, International Journal of Intelligent Systems, pp. 459-468 (2001).

[10] Hart, P.E., The condensed nearest neighbor rule, IEEE Transactions on Informa-tion Theory 14, pp. 515-516 (1968).

[11] Ritter, G.L., Woodruff, H.B., Lowry, S.R., An algorithm for a selective nearest neighbor decision rule, IEEE Transactions on Information Theory 21, pp. 665-669 (1975).

[12] Gates, G., The reduced nearest neighbor rule, IEEE Transactions on Information Theory 18, pp. 431-433 (1972).

[13] Wilson, D., Asymptotic properties of nearest neighbor rules using edited data, IEEE Transactions on Systems, Man, and Cybernetics 2, pp. 408-421 (1972).

[14] Yin-shan Gu,Qiang Hua and Yan Zhan, Case-Base Maintenance Based on Representative Selection for 1-NN, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, vol.4, pp. 2421-2425 (2003).

[15] Yan Zhan, Hao Chen. Reducing Samples Learning for Text Categorization, 3rd International Conference on Information Management, Innovation Manangement and Indeustrial Engineering, pp. 586-589 (2010).