A Review of a Text Classification Technique: K-Nearest Neighbor

R.S. Zhou
The College of Information Engineering
Minzu University of China
Beijing, China

Z.J. Wang
The College of Information Engineering
Minzu University of China
Beijing, China
Minority Languages Branch
National Language Resource Monitoring & Research Center
Beijing, China

Abstract—**In order to get effective information timely and accurately in masses of text, text classification techniques get extensive attention from many aspects. A lot of algorithms were proposed for text classification which made it easy to classify texts, such as Naïve Bayes, Rocchio, Decision Tree, Artificial Neural Networks, VSM, kNN and so on. In this paper, we mainly discussed the latest improved algorithm of kNN including Rocchio-kNN, TW-kNN, RS-kNN and kNN based on K-Medoids. Each of the representative algorithms is discussed in detail. These algorithms based on kNN have reduced the computational complexity as well as increased the execution efficiency compared with the traditional kNN algorithm.**

Keywords-text clasificaton; rocchio-Knn; TW-kNN; RS-kNN; kNN based on K-Medoids

I INTRODUCTION

Text classification is a kind of procedure related with NLP (Natural Language Processing). It finds relational mode (classifier) between text's attributes (feature) and text's category according to a labeled training text corpus, then utilizes the classifier to classify new text corpus. Text classification can be divided into two parts: training and classifying. The purpose of training is to structure classifier which can be used to classify new texts by the connection between training text and category. Classifying means to make the unknown new text assigned with the known category label. The procedure of text classification is showed in Figure.1. (Process I is the procedure of training, Process II is the procedure of classifying.)

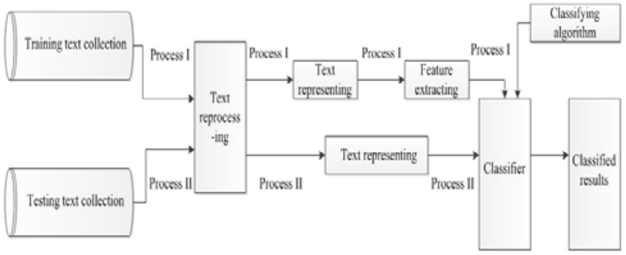

FIGURE I: THE PROCEDURE OF TEXT CLASSIFICATION.

Text reprocessing: it is the first step of text classification which has a great influence on subsequent work. It needs to split sentences into separate words and remove the stop words which have few meaning in the article.

Text representing: natural language texts usually are made up by an unstructured way that can't be recognized by computers. We have to transform this way into another one that can be recognized by computers. This procedure is called text representing. The most widely used way is Vector Space Model (VSM) [1].

Feature extracting: it is the key technique to be resolved in Process I, namely, how to choose those representative features from a large range of features [2].There are some sophisticated methods to extract features, such as IG(Information Gain), MI(Mutual Information), DF (Document Frequency), CHI(χ2-statistics), CC(Correlation Coefficient) [3-7] and so on.

Classifier: Designing classifier is the core of text classification. A classifier's performance is determined by the related algorithms. In this paper, we only discussed the algorithm of kNN and its improved versions.

II KNN ALGORITHM AND ITS IMPROVED VERSIONS

A. *K-Nearest Neighbor (kNN)*

kNN is one of the best algorithms used in VSM, proposed by Cover and Hant in 1967 [8]. Its fundamental idea is based on similarity (a distance function) that measures the distance between untested texts and those known-category texts, then choose the first K texts with least distance. The standard Euclidean distance [9] Sim(di, dj) between document i and document j is often used as the distance function. The Euclidean distance is showed in Eq.1.

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^{2})(\sum_{k=1}^{M} W_{jk}^{2})}} \tag{1}$$

In above equation, Wik and Wjk are the feature's weight of document i and document j. M is the amount of training texts.

The accuracy using kNN to classify texts is much better than most of algorithms, but it has two fatal disadvantages: low efficiency and imbalance. When calculating the similarity, the system has to compare a testing text with all of training texts which leads to low efficiency. Imbalance is caused by the text's amount of different classes in training text. Therefore, the following algorithms of the improvement of kNN are devoted to resolving the disadvantages that can improve the accuracy as well as efficiency in classifying texts.

B. Rocchio-kNN

Rocchio-kNN is proposed by Zheng Z and YongGeng Z in 2004 [10]. Rocchio and kNN are two kinds of common algorithms used in text classification. They are both implemented based on similarity. The advantage of Rocchio is high efficiency but bad results, which is totally opposite compared with kNN. So this algorithm combines Rocchio with kNN to maximize their advantages. Firstly, this algorithm generates candidate of classes for testing texts from training texts with Rocchio [11]. Secondly, choose the final K texts from the candidate with kNN. The rest of work is the same with kNN. Rocchio-kNN utilizes their advantages adequately and avoids their disadvantages as a result of getting Rocchio's efficiency and kNN's good consequence simultaneously. In addition, this approach is easy to be implemented.

C. TW-kNN

TW-kNN is proposed by JinFu Y and Min S in 2011 [12]. The traditional kNN suggests that the contributions made by all of K nearest neighbors are equal, which makes it easy to be disturbed by noises. And as proposed before, the computational demands for classifying texts would be prohibitive. This approach based on Weighted Distance utilizes the template reduction technique to drop samples that are far away from the boundary of classification, which can keep the accuracy of classification and improve the efficiency greatly as the traditional kNN.

The approach makes it more ideal based on TR-kNN [13]. It adopts the template reduction technique to filtrate training texts. After gettingthe new training texts, it will set weight (wt(i)) for K nearest neighbors. Then the class obtaining maximum value of weight according to discriminant function (G(ws, xt)) is selected as the testing text's class. The equation of calculating weight wt(i) is showed in Eq.2. The discriminant function G (ws, xt) is showed in Eq.3.

$$w_{t(i)} = \frac{1}{d_{t(i)}{}^{2}} \qquad \text{i=1, 2, 3……, k} \qquad (2)$$

In above equation, $w_{t(i)}$ is the weight value of No.i. $d_{t(i)}$ is the distance between testing text and the nearest neighbor of No.i.

$$G(w_s, x_t) = \sum_{i=1}^{K} w_{t(i)} I_{t(i)} \qquad \text{s=1,2,3······, c} \qquad (3)$$

Where $I_{t(i)} = \begin{cases} 1, & W_{t(i)} = W_s \\ 0, & W_{t(i)} \neq W_s \end{cases}$ .The maximum G(ws, xt) corresponding to class ws is the class of testing text xt.

D. RS-kNN

RS-kNN is proposed by Ying Y and Duoqian M in 2012 [14]. The classic theory of Rough Set (RS) is a kind of mathematical tool which is used to do quantitative analysis and process inaccurate as well as incomplete information. To make it more resistant to interference, Ziarko proposed a model of Rough Sets based on variable precision [15]. RS-kNN was realized based on the two techniques mentioned above.

This algorithm structures training sample distribution to reduce the amount of computations instead of cutting sample. The purpose of structuring is getting the upper and lower similarity zone for known class utilizing the model of RS based on variable precision. A class's lower similarity zone is its core section, which reflects the class's approximate position, while the element from boundary is uncertain in the upper similarity zone. When classifying a testing text, firstly it judges if this text lying in the center zone of a class, the text belongs to that class if answer is yes. Otherwise, it selects K nearest neighbors from upper similarity zone. Then the left process is the same with kNN.

E. kNN based on K-Medoids

kNN based on K-Medoids is proposed by Xianfeng L and Shenglin Z in 2014 [16]. The effect of this new classifying approach mainly depends on the effect of cluster by K-Medoids whose essence is to choose the center of cluster. K-Medoids technique is used to cluster the training texts into high similarity clusters resulting in high similarity by individuals in the clusters but low similarity among clusters. After finishing cluster, the system needs to produce new training texts set by cut the training texts according to the relative position of the testing texts with the clusters. Lastly, the classifier using kNN can select K nearest neighbors from the new training texts.

K-Medoids is a kind of common cluster algorithm based on partition which has high accuracy and strong robustness in cluster because it's not easy to be impacted by extreme data of sets when existing noise and outliers. However, it still has some disadvantages: 1). Sensitivity in initializing as well as the results of diversification. 2). Low efficiency when rotating center.

III CONCLUSIONS

kNN is a classic algorithm and widely used in text classification. But its two fatal disadvantages make it not so perfect. In this paper, we mainly discussed four kinds of related improvement of kNN in recent years and analyze their operational principle. Whatever version it is, they are all improved from two sides: 1). Reducing amount of training texts by removing those texts with less correlation. 2). Changing the method of choosing K nearest neighbors by combining with other algorithms. In the future, we will do more research to find new improved method to make kNN more efficient and effective.

ACKNOWLEDGEMENTS

In this paper, the research was sponsored by Key Program of National Natural Science Foundation of China (No. 61331013) and National Language Committee of China (No. WT125-46, No. WT125-11).

REFERENCES

[1] Salton G, M E Lesk. Computer evaluation of indexing and text processing. Journal of the ACM, 15(1), pp. 8-36, 1968.

[2] David D. Lewis. Feature selection and feature extraction for text categorization. Proceedings of Speech and Natural Language Workshop, pp. 212-217, 1992.

[3] Lee C K, Lee G G. Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing and Management, 42(1), pp. 55-165, 2006.

[4] R. Nattuti. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks, 5(4), pp. 537-550, 1994.

[5] Salton G, Yang C S, Yu C T. A Theory Term Importance in Automatic Text Analysis. Journal of the American Society for Information Science, 26(1), pp. 33-44, 1975.

[6] Zheng Z H, Srihari S H. Text categorization using modified-CHI feature selection and document term frequencies. ICMLA, pp. 252-263, 2002.

[7] Luigi G, Fabrizio S. Feature Selection and Negative Evidence in Automated Text Categorization. Proceedings of the ACM KDD Workshop on Text Mining, 2003.

[8] Cover T M, Hart P E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp. 21-27. 1967.

[9] PE Danielsson. Euclidean Distance Mapping. Computer Graphics and Image Processing, pp. 227-248, 1980.

[10] Z. Zheng , S.G. Zhou, A.Y Zhou. A New Text Categorization Method Based on kNN and Rocchio. Journal of Computer Research and Development, pp. 226-230, 2004.

[11] J J Rocchio. Relevance feedback in information retrieval. In the SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313-323, 1971.

[12] J.F Yang, M. Song, M.A Li. A Novel Template Reduction K-Nearest Neighbor Classification Method Based on Weighted Distance. Journal of Electronics & Information Technology, 33(10), pp. 2378-2383, 2011.

[13] Fayed H A, Atiya A F. A novel template reduction approach for the k-nearest neighbor method. IEEE Transactions on Neural Networks, 20(5), pp. 890-896, 2009.

[14] Y. Ying, D.Q Miao, C.H Liu, et al. An Improved kNN Algorithm based on Variable Precision Rough Sets. PR & AI, 25(4), pp. 617-623, 2012.

[15] Ziarko W. Variable Precision Rough Sets Model. Journal of Computer and System Science, 46(1), pp. 39-59, 1993.

[16] X.F. Luo, S.L. Zhu, Z.J. Chen, et al. Improved KNN text categorization algorithm based on K-Medoids algorithm. Computer Engineering And Design, 35(11), pp.3864-3867.2014.