BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **16**, No 6
Special issue with selection of extended papers from 6th International Conference on Logistic, Informatics and Service Science LISS'2016

Sofia • 2016 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2016-0078

Chinese Text Auto-Categorization on Petro-Chemical Industrial Processes

Jing Ni[1], *Ge Gao*[2], *Pengyu Chen*[2]

[1]*School of Economics & Management, Beijing Institute of Petrochemical Technology, 102617 Beijing, China*
[2]*MES Department, PCITC Information Technology Co., Ltd., 100007 Beijing, China*
Emails: nijing@bipt.edu.cn gaoge420@bipt.edu.cn pengyu.chen@pcitc.com

Abstract: *There is a huge growth in the amount of documents of corporations in recent years. With this paper we aim to improve classification performance and to support the effective management of massive technical material in the domain-specific field. Taking the field of petro-chemical process as a case, we study in detail the influence of parameters on classification accuracy when using Support Vector Machine* (*SVM*) *and K-Nearest Neighbor* (*KNN*) *Text auto-classification algorithm. Advantages and disadvantages of the two text classification algorithms are presented in the field of petro-chemical processes. Our tests also show that more attention to the professional vocabulary can significantly improve the F1 value of the two algorithms. These results have reference value for the future information classification in related industry fields.*

Keywords: *Text classification; KNN; SVM; petro-chemical; field-specific knowledge.*

1. Introduction

Given the considerable growth of available text documents on the Web and databases, extracting the correct information is crucial for enterprises development [1]. Researchers are highly challenged to find better ways to deal with such huge amount of information [2]. Automatically collecting, processing, sorting and utilizing information can promote the knowledge interaction and knowledge value-added [3]. Text classification aims to classify natural language documents into a fixed number of predefined categories based on their contents [4].

Text classification is an important problem in machine learning, and its research has been attracting attention. However, to the domain-specific text classification problem has not been paid enough attention. At the same time, the demand for text categorization of domain knowledge is increasing. To meet the practical needs, we focus on the domain-specific text classification through the petrochemical processes. Seven classes of petro-chemical processes in petroleum chemical industry are taken with the main concern, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classification methods are compared, and different performance of the two algorithms in the petroleum chemical engineering on Chinese text classification is analysed.

2. Related work

Chinese text automatic classification began in 1981. After 30 years of research, the Chinese text automatic classification technology is becoming mature. The literature on Chinese text classification methods and the research situation are summarized [7-11]. It is claimed that artificial intelligence automatic classification technology and statistical classification techniques are relatively mature. Artificial intelligence automatic classification technology mainly includes establishing the expert system by knowledge engineering techniques to construct knowledge base and set the specific knowledge and experience of experts to pre-established knowledge base, and then constructing the inference engine for reasoning and classification. The technology mainly includes the knowledge base creation and reasoning machine construction. Knowledge expression in the knowledge base mainly includes rule representation, semantic network representation and logical representation; the three kinds of reasoning methods mainly consist of forward reasoning, backward reasoning and hybrid reasoning. Statistics-based classification algorithm is established by comparing similarity between the feature vector and the predefined class feature vector to classify the texts. The technology mainly involves the creation of Thesaurus and Taxonomy, word segmentation method and feature weight vector algorithm. Examples of classification algorithms include Naive Bayes classifier, support vector machines, Kernel estimation (KNN), Neural networks, etc.

2.1. Research on KNN and SVM method

With the rapid development of web technology and digital library, online documents have been rapidly increased, and automatic text categorization has become a key technology to process and organize large amount of document data. As a simple, effective and non-parametric classification method, the KNN method is widely used in text classification. However, KNN classification algorithm efficiency will be greatly reduced when the sample set is getting bigger and becomes high-dimensional data. Therefore, domestic and foreign scholars improved KNN algorithm to improve the efficiency of text classification.

Zhang and Huang [12] gathered samples of numbers of similar documents into core documentation set on text clustering and established a classification model instead of the original sample, which can improve the classification speed. Lu, Zhao and Lin [13] carried out comparative trials of text classification system using KNN method based on discrete value rules and KNN method based on weighted similarity, respectively. Experiments show that the improved algorithm can significantly improve the classification efficiency in the case when the performance of the KNN classification can be kept unchanged; the classification efficiency can be improved significantly [14].

Liu, Yang and Yuan [15] proposed an improved KNN method; compared to the traditional KNN method, the improved KNN method can ensure the accuracy of classification, and the classification efficiency is improved effectively. Due to shortage of KNN in text processing environment, Qian and Wang [16] proposed an improved KNN text classification method based on self-organizing mapping neural network theory, feature selection theory and pattern aggregation theory; it effectively reduced the feature space dimension and improved the speed and accuracy of text classification.

Fan and Chen [17] proposed an improved KNN algorithm based on association analysis. Frequent feature sets for each class of training documents and associated documents should be extracted in advance. When a document with unknown class has to be classified by the use of the results of association analysis, the number of nearest neighbor's k can be decided, k-nearest neighbours can be found quickly from all classes of training documents, and the class of the document can be decided by the classes of its neighbours. Wang et al. [18] introduced the KNN algorithm of central vector classification which has a better classification result compared to the traditional KNN algorithm in processing Chinese text.

For KNN algorithm test complexity is at least linear, resulting in low efficiency in the case of large data samples, a fast KNN classification algorithm is proposed in [19]. The training process is introduced in the k-nearest neighbour algorithm through the linear complexity of clustering method for big data blocking. Then, the algorithm selects the nearest cluster as new training samples and establishes a classification model. This process can greatly reduce the test time of the KNN algorithm, but processing big data causes complexity of the classification.

Experiments of Yang and Liu [20] in the data set Reuters-21578 show that compared with other methods, SVM and KNN method have a certain degree of improvement in recall and accuracy rates. More researchers analyze the accuracy of both SVM and KNN methods in Chinese text categorization. In [21], an empirical study of using the SVM algorithm and the traditional KNN algorithm to categorize the Chinese text is conducted. The experiment shows that compared to KNN, SVM has better categorization effect of the Chinese text and higher recall ratio and pertinence ratio.

Zhang and Pang [22] made a comparison between SVM and KNN on text classification after illustrating the procedures in text classification. The experimental results showed that the accuracy of SVM by using Polynomial kernel function is higher than that by using Radial Basis function. In addition, the accuracy of the former increases when the parameter q gets bigger; the accuracy of SVM and KNN both have higher recall on short texts than on long texts.

A. L. Zhang, G. L. Liu and C. Y. Liu [23] used the SVM algorithm to construct a multi-class classifier which can effectively classify the large scale class into a combination of small class identification problem and can reduce the error recognition rate. The above research only uses the KNN algorithm or the SVM algorithm respectively to carry on the research, without the horizontal comparison about these two kinds of algorithms.

2.2. Fusion of two methods of SVM and KNN

There are advantages and disadvantages in classification by using KNN and SVM methods. Therefore, more researchers combine KNN and SVM algorithms to improve the classification accuracy and the classification efficiency.

Q. Wang, X. L. Wang and Y. Guan [24] proposed a text classification algorithm based on KNN and SVM. KNN algorithm was used firstly to find the class labels, and then a binary SVM classifier was taken to sample the fine points. The experiment is effective at reducing the number of candidate classes and improving the classification accuracy. However, these two classifiers are trained separately by different training samples; it cannot be guaranteed that the classification facet in the classification output had good comparability. On the other hand, the assumption of the single layer text class structure in practice was uncommon.

In [25], the improved KNN-SVM that combined SVM with KNN is presented to improve the accuracy of imbalanced classification nearby SVM hyper plane. A large amount of experiments by the UCI dataset show that the algorithm can significantly improve the identification rate of the minority samples and overall classification performance. Based on SVM and KNN algorithms, Kuang and Xia [26] put forward a SVM-KNN algorithm combining KNN and SVM classifiers. Test shows that the SVM-KNN algorithm can improve the performance of classifier by feedback and improvement of classifying prediction probability.

2.3. Application of text classification technology

Classification methods have been applied in various fields for retrieval [27-30], spam filtering [31-32], authors identification [33], web page classification [34], and emotional analysis [35].

Li [36] used SVM to classify the web text and compare the results with KNN, the advantages of SVM algorithm in efficiency are shown; Chen and Fu [37] used SVM in emotion recognition and text classification, and constructed the sentiment polarity classification model; Through the SVM active learning algorithm. Gao et al. [38] studied the classification of four different levels of We Chat articles using the information of alarm degree evaluation analysis technology, and constructed We Chat information early warning system. In Chen and Fu [39], the SVM classification algorithm is used to generate the micro-blog information classification results, which include a large number of tourism related concepts of vocabulary, as the construction and expansion of the corpus of tourism ontology. Chen and Tang [40] used SVM to classify the sentiment, and then extracted semantic terms from the comments through the LDA model. Deng and Wu [41] classify the topic of the web site by class center vector algorithm and SVM. The method was applied to the forestry business information resource, and the effectiveness of the method is verified. Jiang and Sun [42] used the SVM model to study the automatic extraction of terms in Chinese domain.

With the continuous progress of science and technology, information technology has been developed further. Academic papers, technical reports, technology blogs, professional books and other academic text resources update quickly. Having continuous accumulation of quantity and categories, using text categorization methods effectively becomes a challenging research topic in information field for extraction of the required information from a large amount of information.

3. Experimental section

3.1. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) algorithm was introduced by Dasarathy [5] in 1991. For KNN, each document is assigned to the majority class of its k closest neighbors where k is a parameter. The rationale of KNN classification is that, based on the contiguity hypothesis, we expect a test document d to have the same label as the training documents located in the local region surrounding [6].

For KNN algorithm, the selected neighbors are objects which have been correctly classified. This method determines the categories of the samples to be classified according to the closest one or a few samples in the decision making. Due to KNN, the method mainly rely on the proximity of limited samples around, rather than on distinguishing the domain classes to determine the category, so for crossing or overlapping domain sample set to be divided, KNN method is more suitable than other methods.

There exists a set of labelled training sets; when unlabelled new data is entered, each feature of new data is compared to corresponding feature of labeled training set, then labels of the most similar (nearest neighbour) category in samples are extracted. In order to classify a testing article, it computes the distance between this article and all the training articles. Then, the *k* training articles of closest distance to the test article are used to select the word features. The algorithm is described as follows [6]:

```
TRAIN-KNN(¢, D)
Step 1. D'←PREPROCESS(D)
Step 2. k←SELECT-K(¢, D')
Step 3. return D', k
APPLY-KNN(¢, D', k, d)
Step 1. Sk←COMPUTE NEAREST NEIGHBORS(D', k, d)
Step 2. for each cj ∈¢
Step 3. do pj←|Sk∩cj |/k
Step 4. return argmaxj pj
```

Where

¢ is Set {*c*1, . . . , *cJ*} of all classes

C is random variable that takes as values members of *¢*

D is Set {⟨*d*1, *c*1⟩, . . . , ⟨*dN*, *cN*⟩} of all labeled documents

Sk(*d*) is the set of *d*'s k-nearest neighbors *pj* is an estimate for *P*(*cj* |*Sk*) = *P*(*cj* |*d*); *cj* denotes the set of all documents in the class *cj*.

The parameter *k* in KNN is often chosen based on experience or knowledge about the classification problem at hand.

3.2. Support Vector Machine (SVM)

SVM were introduced by Vladimir Vapnik [22]; they are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on that side of the gap where they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [45].

3.3. Chinese document categorization framework

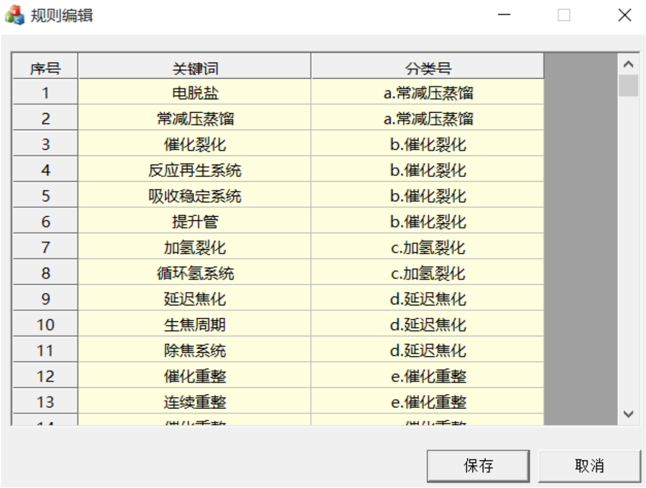
Text classification generally includes text pre-processing, statistical calculation, feature selection, classification algorithm selection and effect evaluation.

3.3.1. Text pre-processing

Document pre-processing is the first step in text classification, which converts the Chinese documents to a form that can meet the needs of classification. These preprocessing tasks include text tokenization, stop word removal, and word frequency statistics.

Considering each document may be filed to more than one category. If the general dictionary is only used to preprocess the text in specific field, it is obviously that errors will directly affect the accuracy of text classification in the segmentation stage, because domain-specific terms tend to be wrongly separated. Furthermore, domain-specific texts tend to have many new terms. In petrochemical field, more and more new synthesis technology and process route appear with the advancement of technology and continuous innovation. In addition, professional texts often present several professional symbols, professional name abbreviations, etc., which need special consideration during the pre-classification. Therefore, we combine the domain knowledge with the general dictionary to improve the efficiency of domain text classification.

Chinese text tokenization is to separate continuous text content into word sets (bag of words), considering the problem of Chinese “three noes”. We use stop words dictionary to support elimination of insignificant words, such as “的、地、得、着、了、过”. These tokens could be individual words that are converted without understanding of their meanings or relationships. The list of tokens becomes input for further processing. Then we use professional domain dictionary to avoid a large number of professional vocabulary to be wrongly divided into meaningless single words. Before that, mapping relationship between specific keywords, words and classification number are added it to professional dictionary, as shown in Fig. 1. At last, IKAnalyzer custom word segmentation dictionary is used to get term and its number of times.



序号	关键词	分类号
1	电脱盐	a.常减压蒸馏
2	常减压蒸馏	a.常减压蒸馏
3	催化裂化	b.催化裂化
4	反应再生系统	b.催化裂化
5	吸收稳定系统	b.催化裂化
6	提升管	b.催化裂化
7	加氢裂化	c.加氢裂化
8	循环氢系统	c.加氢裂化
9	延迟焦化	d.延迟焦化
10	生焦周期	d.延迟焦化
11	除焦系统	d.延迟焦化
12	催化重整	e.催化重整
13	连续重整	e.催化重整

Fig. 1. Domain professional vocabulary rule editor

3.3.2. Statistical calculation and feature selection

We adopted TF-IDF statistical algorithm [44] to weight. The basic idea is that the weight of a phrase in a document vector is the product of local and global parameters. If a word or phrase in a single document term frequency TF is high, but rarely in other documents, the word or phrase is considered to be strong for class distinguishes between attributes.

Generally speaking, the feature space dimension of structured text is high and needs to be reduced and only retains some of those close to the content. During the feature selection, priority was given to the terms in professional vocabulary. After calculating document feature weight TF-IDF vector (term frequency inverse document frequency) to obtain the statistical information, feature terms in the specialized dictionaries should have priority in order to be selected as category terms; then the TF-IDF values are sorted in descending order; finally, the top N entries are taken as a feature vector for text classification.

3.3.3. Classifier performance evaluation

The evaluation of the performance for classification model to classify documents into the correct category is conducted by using several mathematic rules such as recall, precision, and *F*-measure, which are defined as follows:

(1) $$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}},$$

(2) $$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}},$$

where TP is the number of documents that are correctly assigned to the category, FP is the number of documents incorrectly assigned to the category, and FN is the number of documents that belonged to the category, but are not assigned to the category. The success measure, namely, micro-*F*1 score, a well-known *F*1 measure, is selected for this study and is calculated as follows:

(3) $$\text{F1} = \frac{2^*\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}.$$

3.3.4. Test dataset

Training set involves 7640 documents and 5546 test documents. According to the process of petrochemical processing, the documents fall into seven categories: atmospheric-vacuum distillation, catalytic cracking, catalytic reforming, hydrogenation refining, hydrogenation cracking, delay coking, and hydrogenation residue.

Table1. Test dataset

No	Category	Training set	Test set
1	atmospheric-vacuum distillation	1500	1000
2	catalytic cracking	1500	1000
3	catalytic reforming	1000	810
4	hydrogenation refining	1300	1033
5	hydrogenation cracking	1200	996
6	delay coking	1060	567
7	hydrogenation residue	780	143

4. Results and discussion

4.1. KNN parameter optimization

In the KNN algorithm, *k* value is one of the most important parameters, especially for the text classification with strong specialization. In the early stage of testing, the value of the *k* parameter of the KNN should be optimized. If *k* value is too low, the number of neighbours is rare, which will reduce the classification accuracy. If *k* value is too high, it is easy to increase the noisy data and reduce the accuracy of classification. The number of valid parameters is related to the *k* value, roughly equal to *n/k*, where n is the number of documents in this training data set. The *k* value is determined by several experiments of determination according to minimum classification error rate.

The relation between *k* values and *F*1 based on categories of "atmospheric-vacuum distillation" and "catalytic cracking" is shown in Fig. 2. When value of *k* is increased, the classification accuracy is overall downward. With the purpose of a high classification accuracy rate, *k* value should be selected as low as possible. We tested k value from 5 to 35 and compare the influence *F*1 vs *k*; we selected *k*=15 to the KNN algorithm.

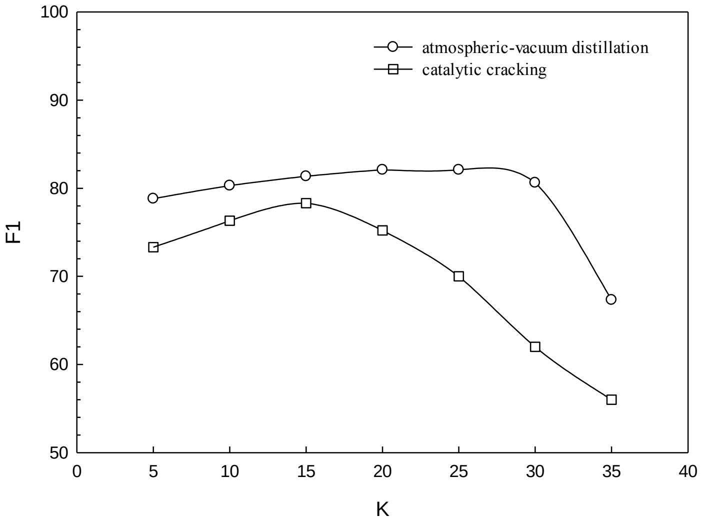


Fig. 2. Influence of K on text classification in KNN algorithm

4.2. Parameters determination of SVM Algorithm

The key step of the SVM algorithm is to select or construct the kernel function. Samples can be transformed to the kernel function matrix by kernel function, equivalent to mapping the input data to high-dimensional feature space through a nonlinear function. In the feature space, kernel function matrix is converted into nonlinear model of the input space by various linear algorithms.

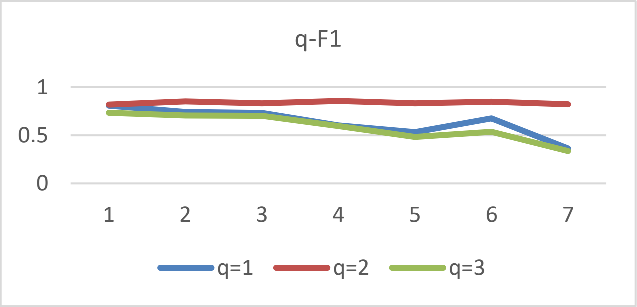
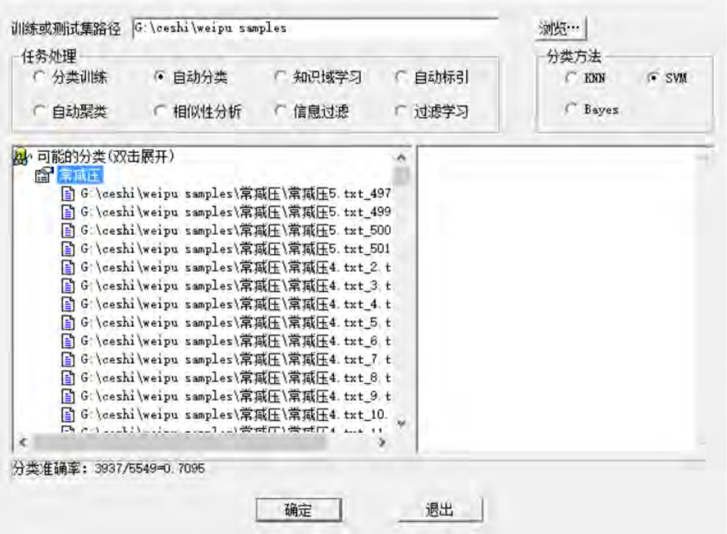


Fig. 3. SVM parameter optimization & classification result

SVM common kernel function is polynomial kernel function and radial basis kernel function; the study shows that the classification accuracy of using polynomial kernel function is higher than that of the radial basis kernel function [22]. The form of polynomial kernel function is $K(x_i, x_j) = (x_i \cdot x_j + 1)^q$, Polynomial kernel function has only one parameter q. When using different q, the performance of the text classification fluctuated, and the required training time is increased with the increase of the parameter q. In this paper, we select 1, 2, 3 for the q value respectively when optimizing the *q* value, and the effect of the classification results on *F*1 is tested as shown on Fig. 3.

It can be seen on the Fig. 3 that *F*1 decreases with the increase of *q* value, so *q* value in the petrochemical processing text classification is selected as 2.



训练或测试集路径 G:\ceshi\weipu samples

任务处理
○ 分类训练 ● 自动分类 ○ 知识域学习 ○ 自动标引
○ 自动聚类 ○ 相似性分析 ○ 信息过滤 ○ 过滤学习

分类方法
○ KNN ● SVM
○ Bayes

```
可能的分类(双击展开)
  常减压
    G:\ceshi\weipu samples\常减压\常减压5.txt_497
    G:\ceshi\weipu samples\常减压\常减压5.txt_499
    G:\ceshi\weipu samples\常减压\常减压5.txt_500
    G:\ceshi\weipu samples\常减压\常减压5.txt_501
    G:\ceshi\weipu samples\常减压\常减压4.txt_2.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_3.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_4.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_5.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_6.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_7.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_8.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_9.t
    G:\ceshi\weipu samples\常减压\常减压4.txt_10.
```

分类准确率：3937/5549=0.7095

Fig. 4. SVM classification result

4.3. Comparison between KNN and SVM

In this test, the documents in each category were first pre-processed by converting them to UTF8 encoding. For stop words removal we used stop list containing 1208 stop words.

Feature selection was performed using TF-IDF values. The numbers of features for constructing the vectors were 100, 300, and 500. In doing so, the effect of pre-processing task can be comparatively observed within a wide range of feature size. Feature vectors were built using the function TF-IDF. After the selection of optimal *k*=15 of the KNN algorithm and *q*=2 condition to SVM algorithm, the seven petrochemical processes are compared. Fig. 4 presents the classifier interface, and the classification results are shown on Figs 5-7. Table 2 shows Cross-references between numbers and class names.

Table 2. Cross-references between numbers and class names

Number	Class name
1	atmospheric-vacuum distillation
2	catalytic cracking
3	catalytic reforming
4	hydrogenation refining
5	hydrogenation cracking
6	delay coking
7	hydrogenation residue

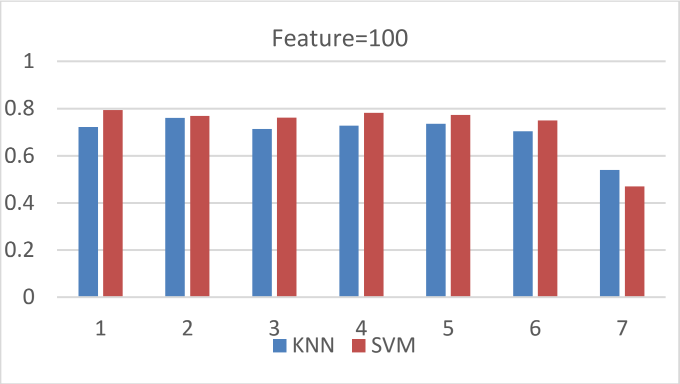


Fig. 5. *F*1 for both KNN and SVM with number of feature=100 for each category

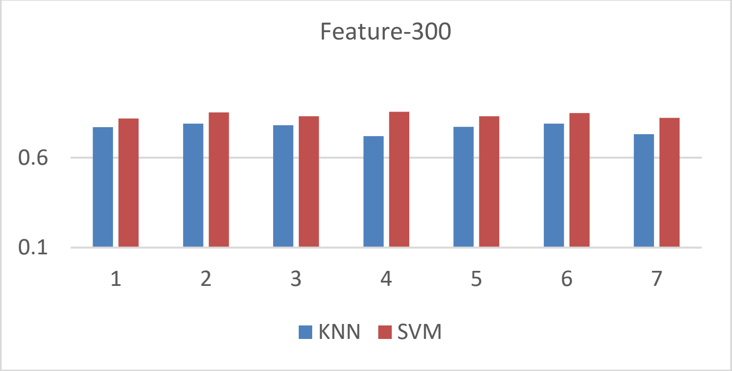


Fig. 6. *F*1 for both KNN and SVM with number of feature=300 for each category

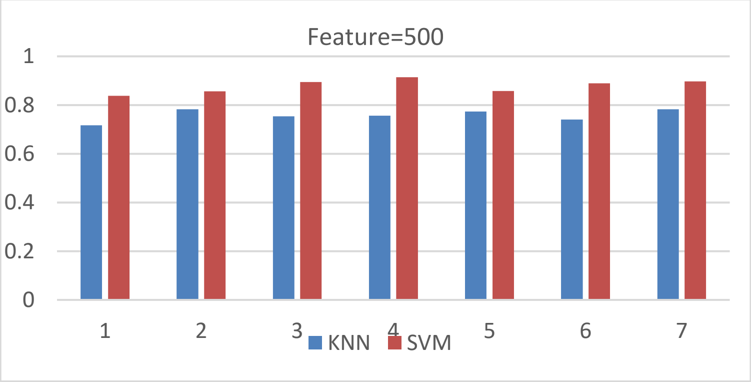


Fig. 7. *F*1 for both KNN and SVM with number of feature=500 for each category

The above figure depicts the *F*1 values for varying numbers of features from 100 up to 500. When the number of features is small, KNN is only a little inferior to SVM. As the number of features increases, the SVM outperforms the KNN. This is due to the clearer optimal hyperplane obtained by the increasing number of features and has the ability to handle high dimensional data, whereas the KNN performance degrades slightly. This is because KNN is a lazy algorithm that depends barely on statistics and comparison, and must keep a track of large amount of features.

The result shows that even if the feature words in the field of professional field are added, the involved field concepts are too much, which will lead to the decrease of the professional division degree. Therefore, the manual intervention of fuzzy terms is necessary.

5. Conclusion

To answer the demand of scientific and technical personnel and information personnel to classify texts, this paper presents a study on the automatic classification of knowledge in the field of petro-chemical industry. On the basis of comparing the performance of SVM and KNN algorithms in Chinese text classification, it is proved that the SVM classifier is superior to KNN. The results clearly showed the superiority of the SVM over the KNN algorithms in all experiments.

Findings of this study showed that following points should also be considered:

(1) The usage of domain knowledge dictionary in the text pre-processing stage strongly affects the text categorization result. If the text contains no professional vocabulary and only common sense knowledge base words are used, then a large number of professional vocabularies will be divided into meaningless single words. Classification performance may significantly suffer.

(2) Some of texts have low frequency field terms. To avoid being filtered in the feature ranking algorithm, feature amplification of the vocabulary in the field value is needed.

(3) In order to better reflect the important role of the field vocabulary in the classification, the weight of the vocabulary is calculated according to a certain proportion to distinguish chemical industry standard, chemical industry field vocabulary and general vocabulary.

Further research can be undertaken to develop an ontology application for the rule edition of petro-chemical product, which can be embedded in our classification system.

***Acknowledgements*:** This work is sponsored by the Education Ministry Funding Project for the Development of Liberal Arts and Social Sciences (No 12YJA87001, 15YJAZH052), Beijing Social Science Fund (No 14SHB010), National Training Program of Innovation and Entrepreneurship for Undergraduates (No 2016J00134).

References

1. Ni, J., X. Li, J. Zhou et al. Ontology-Based Knowledge Modelling and Reasoning for Petrochemical Products. – China CIO News, Vol. **7**, 2012, No 20, pp. 138-140.
2. Abdullah, A., G. Tan, A. Khaled et al. The Effect of Preprocessing on Arabic Document Categorization. – Algorithms, Vol. **9**, 2016, No 2, p. 27.
3. Xiao, L., J. L. Huang. The Application of Knowledge Management in Petroleum and Petrochemical Industry. – Oil Forum, Vol. **5**, 2014, pp. 29-35.

4. Thorsten, J. Text Categorization with Support Vector Machines. Learning with Many Relevant Features. Berlin, Springer, 1998.
5. Dasarathy, B. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. – In: McGraw-Hill Computer Science Series. Las Alamitos, California, IEEE Computer Society Press, 1991.
6. Manning, C. D., P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press, UK, 2009
7. Wang, R. Q., T. J. Hu. Automatic Classification of Chinese Text-State of Art. – Journal of Medical Intelligence, 2002, No 6, pp. 342-347.
8. Niu, Y. L., H. Zhang. Automatic Classification of Chinese Text-State of Art. – Software Guide, Vol. **7**, 2008, No 4, pp. 24-26.
9. Tan, J. B., Y. Li, X. J. Yang. Development of Text Automatic Categorization Measurement Research. – New Technology of Library and Information Service, 2005, No 5, pp. 46-49.
10. Yuan, J. P., D. H. Zhu, Y. Li. Survey of Text Mining Technology. – Application Research of Computer, 2006, No 2, pp. 1-4.
11. Zhan, Y., H. Chen, F. Yuan, L. J. Wang. Research Progress of Text Mining Technology. – Journal of Hebei University (Natural Science Edition), Vol. **23**, 2003, No 2, pp. 221-226.
12. Zhang, X. F., H. Y. Huang. An Improved KNN Text Categorization Algorithm by Adopting Cluster Technology. – Pattern Recognition and Artificial Intelligence, Vol. **22**, 2009, No 6, pp. 936-940.
13. Lu, Z. Y., S. Zhao, Y. M. Lin. Research of KNN in Text Categorization. – Computer and Modernization, Vol. **11**, 2008, pp. 69-72.
14. Sun, R. Z. An Improved KNN Algorithm for Text Classification. – Computer Knowledge and Technology, Vol. **6**, 2010, No 1, pp. 174-175.
15. Liu, B., L. Yang, F. Yuan. Study on Intelligent Question-Answering System of Restricted Field. – Journal of Xihua University (Natural Science), Vol. **27**, 2008, No 2, pp. 33-36.
16. Qian, X. D., Z. O. Wang. Text Categorization Method Based on Improved KNN. – Information Science, Vol. **23**, 2005, No 4, pp. 550-554.
17. Fan, H. L., W. Q. Chen. An Improved KNN Approach of Text Classification Based on Association Analysis. – Computer Technology and Development, Vol. **24**, 2014, No 6, pp. 71-74.
18. Wang, A. P., X. Y. Xu, W. W. Guo et al. Text Categorization Method Based on Improved KNN Algorithm. – Software Technology, Vol. **30**, 2011, No 18, pp. 8-10.
19. Su, Y. J., Z. Y. Deng, D. B. Chen. Fast KNN Classification Algorithm under Large Data. – Application Research of Computers, Vol. **33**, 2015, pp.1-6.
20. Yang, Y., X. Liu. A Re-Examination of Text Categorization Methods. – In: Proc. of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, ACM, pp. 42-49.
21. Liu, H. L., Z. G. Zhang, Z. H. Ma et al. An Empirical Study of Chinese Text Categorization Based on SVM and KNN. – Information Studies: Theory and Application, Vol. **31**, 2008, No 6, pp. 941-944.
22. Zhang, H. X., J. G. Pang. Research on Text Classification Based on SVM and KNN. – Journal of Modern Information, Vol. **35**, 2015, No 5, pp. 73-77.
23. Zhang, A. L., G. L. Liu, C. Y. Liu. Research on Multi-Class Text Categorization Based on SVM. – Journal of Information, 2004, No 9, pp. 6-10.
24. Wang, Q., X. L. Wang, Y. Guan. A Research on Text Categorization Based on Fusion of KNN and SVM. – Chinese High Technology Letters, Vol. **5**, 2005, No 15, pp. 19-24.
25. Wang, C. X., T. Zhang, C. S. Ma. Improved SVM-KNN Algorithm for Imbalanced Datasets Classification. – Computer Engineering and Applications, Vol. **52**, 2016, No 4, pp. 51-55.
26. Kuang, C. L., Q. Q. Xia. Analysis on Text Classification Algorithm Based on SVM-KNN. – Computer Era, 2010, No 8, pp. 29-31.
27. Cateni, S., V. Colla, M. Vannucci. Variable Selection Through Genetic Algorithms for Classification Purposes. – In: Proc. of 10th IASTED International Conference on Artificial Intelligence and Applications (AIA'10), 2010, pp. 6-11.

28. Yang, T., X. J. Lu, Z. F. Liao, W. Liu, H. A. Wang. SVM Based Decision Analysis and Its Granular-Based Solving, – Lecture Notes in Computer Science, Vol. **5593**, 2009, pp. 734-748.
29. Tsai, C. F., Z. Y. Chen, S. W. Ke. Evolutionary Instance Selection for Text Classification. – The Journal of Systems and Software, Vol. **90**, 2014, pp. 104-113.
30. Ghiassi, M., M. Olschimke, B. Moon, P. Arnaudo. Automated Text Classification Using a Dynamic Artificial Neural Network Model. – Expert Systems with Applications, Vol. **39**, 2014, pp. 10967-10976.
31. Gunal, S., S. Ergin, M. B. Gulmezoglu, O. N. Gerek. On Feature Extraction for Spam e-Mail Detection. – Lecture Notes in Computer Science, Vol. **4105**, 2006, pp. 635-642.
32. Uysal, A. K., S. Gunal. A Novel Probabilistic Feature Selection Method for Text Classification. – Knowledge-Based Systems, Vol. **36**, 2012, pp. 226-235.
33. Cheng, N., R. Chandramouli, K. P. Subbalakshmi. Author Gender Identification from Text. – Digital Investigation, Vol. **8**, 2011, pp. 78-88.
34. Ozel, S. A. A Web Page Classification System Based on a Genetic Algorithm Using Tagged-Terms as Features. - Expert Systems with Applications, Vol. **38**, 2011, pp. 3407-3415.
35. Maks, I., P. Vossen. A. Lexicon Model for Deep Sentiment Analysis and Opinion Mining Applications. – Decision Support Systems, Vol. **53**, 2012, pp. 680-688.
36. Li, Q. Text Classification Based on SVM Network. – Electronic Technology, Vol. **10**, 2014, pp. 8-11.
37. Chen, P. W., X. F. Fu. Research on Sentiment Classification of Text Based on SVM. – Journal of Guangdong University of Technology, Vol. **31**, 2014, No 3, pp. 95-101.
38. Gao, X. W., S. Y. Zheng, L. Gao et al. WeChat Monitoring Research Based on SVM Active Learning. – Computer & Digital Engineering, Vol. **44**, 2016, No 4, pp. 715-719.
39. Li, Z. Y., X. W. Yang, M. Wang. Research on the Classification of Travel Demand Information and the Acquisition of Ontology Concept Based on "We Media". – Library and Information Service, 2015, No 23, pp. 106-114.
40. Chen, G. L., W. M. Tang. Text Topic Mining Based on Sentiment Classification. – Journal of Chongqing Normal University (Natural Science), 2016, No 1, pp. 92-96.
41. Deng, H. P., G. Wu. Discovery of Topic-Specific Information Source Based on Web Crawler and Website Classification. – Computer Engineering and Applications, Vol. **52**, 2016, No 3, pp. 59-65.
42. Jiang, T., J. J. Sun. Automatic Chinese Field Terminology Extraction Based on SVR Model. – Information Studies: Theory & Application, Vol. **39**, 2016, No 1, pp. 24-31.
43. Yang, L. H., Q. Dai, Y. J. Guo. Study on KNN Text Categorization Algorithm. – Microcomputer Information, Vol. **22**, 2006, No 7, pp. 269-271.
44. Salton, G., A. Wong, C. S. Yang. A Vector Space Model for Automatic Indexing. – Common ACM, Vol. **18**, 1975, pp. 613-620.
45. Support vector machine. **https://en.wikipedia.org/wiki/Support_vector_machine**