PROCEEDINGS OF SCIENCE

The Mixture of Pattern Aggregation and CHI Statistical Techniques of the Categorization Research in Uyghur

Qiang Ma[1], **Ailin Li**[2], **Hongzhi Yu**[3]
Key Laboratory of National language intelligent processing
National Languages Information Technology Northwest University for Nationalities
Gansu, Lanzhou 730030, China
Email: lalwxy@163.com

Feature selection in pattern recognition, data mining, and other fields has a very wide application. Compared with a variety of feature selection methods in the practical application of the Uyghur text categorization, this paper uses CHI statistical method for feature selection, as traditional CHI statistic is calculated only showing the entry of contribution degree, but not explaining entries and the categories of correlation, and therefore, this paper improves the CHI statistic, and combines with pattern aggregation method, so as to realize the Uyghur text feature selection. Experimental results show that the proposed approach in this paper can not only save the time of feature selection and classification, but also greatly improve the quality of the classification.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1] Speaker

[2] This study is supported by Key Laboratory of National language intelligent processing, Gansu Province (Northwest University for Nationalities)

[3] Correspongding Author

© *Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).*

1. Introduction

With the advent of the era of big data, vast amounts of data information fills in the internet where multilingual information resources show a trend of exponential growth, including Uyghur. The text is the most basic carrier of the data information, so how to effectively manage and use these numerous and complicated information has become a focus in research between Internet enterprises and research institutions. Text classification is the key technology of acquisition and organizes a large number of text data, and it's widely used in information retrieval, text filtering, data mining,resources recommended and other fields[1].

Automatic text categorization is a procession that an unknown categoriy, according to its content, will automatically become one or more categories under the given text. The main procession of text classification includes text preprocessing, feature dimension reduction, classified building (training and texting), and the structure of the classification evaluation. Every link of classification will affect the final result. The quality of the feature dimension reduction directly affects the cost, speed, accuracy of the classification task. In recent years, the study of feature extracting in text categorization has become white-hot. There are two methods of feature dimension reduction: one is the feature selection, and the other is feature extracting. Feature selection refers to the subset that is selected from the characteristic collection, which is useful, and consistes of categories. Feature extraction is mainly used to solve the problem of large number of synonyms and ambiguous words in natural language. At present, the text classification adopts a variety of feature selection algorithms, such as information gain, CHI statistics, mutual information, expected cross entropy, document frequency and other methods[2].

The study about the method of the text feature selection in Chinese is relatively larger, in 2009, He Liu, puts forward a feature weighting method of text classification based on the importance of characteristics, this method is based on rough set theory, by defining the importance of characteristics, and it introduces the decision information leading from characteristics to the feature weighting[3]. Nameng Jia, in 2011, put forward a method of feature selection called FCD, which is the ability based on the contribution of characteristics. This method regards the contribution as the condition that is selected[4]. For Uyghur, at present, most of the research only used items to extract features. In 2011, Alimjan AYSA and others are used by the method of combination of stem segmentation and information gain to express the reduction of space dimension[5].

2.Text Preprocessing

Text preprocessing is formatting primitive data in order to facilitate subsequent processing of the text. For text classification problems, text pretreatment means to format unification while the procession is to extract the keywords of the represent text information.

The text pretreatment procession includes: getting and denoising the original corpus and stem segmentation, and removing the stop words (Meaningless words), and text representation.

2.1Accessing and Denoising the Original Data

PoS(ISCC2015)033

This paper mainly receives from the People's Daily Online in Uyghur (http://uyghur.people.com. cn/), Tianshan net in Uyghur (http://uy.ts.cn/), and so on. Owing to the rich full web tags contained in different kinds of Internet, the text categorization is just with for the content, therefore, the procession of denoising the corpus which comes from Internet is complex.

2.2 Stem segmentation

To such kind of adhesion of language in Uyghur, word obtaining is quite simple, because the documents of these languages themselves are separated by spaces of punctuation, word segmentation is not the key point. The characteristics of agglutinative languages are those that have more complex temporal change and abundant morphological structure, so the difficulty is the segmentation of stem, which means that extract the parts of the meaningful words and regard them as feature items.

This paper, according to the existing dictionary and the morphological rules in Uyghur and other available resources, adopts the mixed processing method of combination of rule and dictionary to achieve the most basic stem segmentation, analyzes from A to B, and uses the algorithm of Lovin to search the match affix from dictionary, out of word affix.

2.3 Building a Stop Word-list

In order to achieve the goal of high efficiency and low consumption, we often filter out some words from the text, which are called the stop words. They're the same as the task of text classification, for a document, some words in the document having no contribution for classification work, such as prepositions, conjunctions, and so on. Filtering out these words can reduce the complexity of the classification task. In this paper, we use two methods to build stop word-list: artificial structures and automatic learning based on statistics.

2.4 Text Representation

Before classifying the text in Uyghur, we need to represent the text in the form that the computer processes. Vector space model, is one of the methods that is widely used and has better effect. Vector space model is a statistical model about the document representation. The text is expressed as feature characteristics of weigh vector, at a point in the n-dimensitional space to show a document. In Uyghur text, features can choose words, phrases or word u-gram to represent the vectors of each component.

In the VSM, each document is represented as the following forms of vectors: Di= (ti1, wi1; ti2, wi2; .... tin, win), tik represents feature item, and wik represents the weight.

3. A Feature Selection Method Based on Improved x^2 and Pattern Aggregation

There are 32 letters in Uyghur, each of them has four different forms. Several words in the text evolve from the same root, as grammatical variants with each other. Words morphologically change, but the meaning has no big difference. Reflected in large amount of dimension of the original feature space, the article has more sparse representation and more flexible morphological changes.

After using stem segmentation and stop-word wiping, although the dimensions of the text of the feature space are still ten thousand, which for most of the classifications is intolerable, and such High Dimensional Characteristic of the classification procession may not be important, even can interfere the training effect, or greatly reduce the classification performance, it is still necessary to take measures to further reduce the dimensions of the feature space.

This paper proposes an improved x^2 to measure the entry's contribution to the text classification. Then according to the model aggregation theory, it makes the similarity entries that has contribution to the text into the same feature item, and builds a feature vector space model of text.

This method can effectively reduce the dimension of vector space. Feature extracting selected from text features is the most representative part, to reduce the dimensions of feature space, and to reduce the computational complexity and improve the accuracy of classification. As a result, the feature vector model should mostly reflect the content of the text. In this paper, the so-called model refers to one of the characteristics of a feature space which is one dimensional feature space. Before aggregating, each pattern corresponds to only one entry, known as the primary model.

For the feature items that are similar in the ratio of contribution of classification, though with different weights, they have the same effect on classified operation, so the classified operation can be regarded as the same model. For this reason, according to the theory of model aggregation that is similar in the ratio of contribution of classification, it will merge into one pattern. This work is called pattern aggregation. The polymerization of a new model will contain one or more characteristics, which greatly reduce the dimension of text vector.

After evaluating every entry's contribution to each category, which based on the combination of the aggregation method and the classified contribution which has similar ratio, at last we get a text characteristic with lower dimension.

3.1Improvement of x^2

The measure of x^2 is a correlation between a feature item and a category. The greater values of x^2 , the closer correlation between the items and category which the items belong to it, that is to say, the item has a lot of information, and x^2 also thought the state of feature existing and unexisting. If there is a feature T and category C, the computation formula of x^2 is as follow:

$$x^2(t,c)=\frac{N\times(A\times B-B\times C)}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{3.1}$$

Which, A represents the occurrences of category C and characteristic T appearing at the same time. B represents both category C and characteristic T do not appear. N represents the total number of documents. It is important to note that when using CHI, we usually need to set a threshold that is a reference point, and regard the feature items which are higher than the threshold as a candidate.

This traditional calculation method of x^2 just show entry contribution to the categories, but doesn't tell the relevance between entry and category. Because the entry and category correlation has both positive and negative cases, when $A\times B-B\times C>0$,it means they have

PoS(ISC2015)033

positive relationship, also means items appear, and some category may appear, the greater amount of x^2, the more possibility that some items belong to some category in the text. On the other hand, when $A\times B-B\times C<0$ they have negative relationship, and items' appearance may not explain the category's appearance. So even if entries have the same amount to some, for x^2 the text, they are vice versa. The computation formula of improvement of x^2 is as follow:

$$x^2(t,c)=sign(A\times D-B\times C)\frac{N\times(A\times B-B\times C)}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{3.2}$$

Which

$sign(x)=1, x\geqslant 0 ; sing(x)=-1, x<0$

The improved x^2 not only show the entry contribution to a classification of a text category, but also show the correlation between the entry and the text, that is, the value of sign () is 1, and the text contains the possibility of a certain category of the entry, if the value of sign () is -1, the entry appears, and the corresponding category is likely not to appear. Generally characteristics' CHI values are the average or maximum to all kinds of x^2. In this paper, on the improvement of x^2, it stipulates the value of entry and CHI:

$$CHI_i=max\left\{x\frac{2}{i1}, x\frac{2}{i2}, \ldots, x\frac{2}{in}\right\} \tag{3.3}$$

3.2 The Feature Dimension Reduction Based on the Theory of Model Aggregation

The reduction procedure of feature dimension that based on improved x^2 and the method of the model aggregation are as follows:

(1)according to the formula (2), calculate the x^2 of each entry for each type;

(2)according to formula(3), calculate each entry's CHI, then order them from high to low, based on the characteristics, and select several items that are in front of the larger CHI, to receive a matrix that has the number of M for models;

(3)compare whether the classification that every model gives to every classification has the same value or not. First of all, every model's improved x^2 between [-1,1], the method is just as the formula:

$$A_{ij}=x\frac{2}{ij}/(max-min) \tag{3.4}$$

Among them, max, min separately represent the maximum and minimum of the model's improved x^2 ;

(4)use the simplest clustering algorithm, based on model A, and each line of A represents a pattern. The same pattern of polymerization is a new model that gets the number of L for it, which is far less than the number of M.

This paper uses the rank of condensation method for clustering, distance between measurement is adopting the most commonly measurement called EUCLIDEAN DISTANCE, look at the formula:

$$d(i, j)=\sqrt{(A_{i1}-A_{j1})^2+(A_{i2}-A_{j2})^2+\ldots+(A_{in}-A_{jn})^2} \tag{3.5}$$

PoS(ISCC2015)033

The EUCLIDEAN DISTANCE is less than a certain threshold model for clustering, the procession is:

① according to matrix A, calculate the model whose distance is less than threshold and cluster it.;

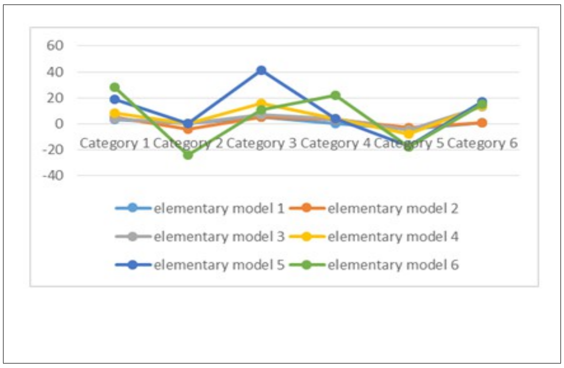
② after clustering, each type of the pattern is combined into another pattern, and this pattern includes all the entries of the class. The word frequency is the sum of these terms. Recalculate the new model of the improved x^2 , according to the new mode to form matrix A;

Repeat these two steps, until all the models can't be clustered;

(5)recount every feature's CHI, according to CHI to choose the number of I for feature.

3.3 Illustration

Hypothetical category equals 6, there are six primary models, according to the date from Fig.1, that is the distribution curve of classified contribution for primary model based on improved x^2 , after the standard processing in it, you can get a new one in the second figure. In Fig.2, some models of the curve for classified contribution become very close after the standardized procession, such as model 1, 4 and 6 e similar, and model 2 and 5 are similar. As a result, these two groups of models can be aggregated. Therefore, after polymerization six primary models have changed to three models, pattern 1 includes elementary model 1, 4 and 6, pattern 2 includes elementary model 2 and 5, and pattern 3 includes model 3.



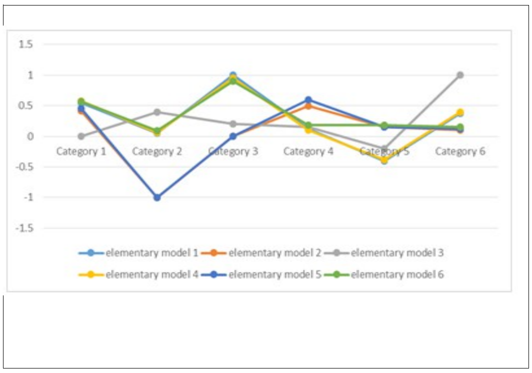


Figure 1: The Distribution Curve of Classified Contribution for Primary Model

Figure 2: The Distribution Curve of Classified Contribution after Standardization

4. Support Vector Machine Classification Algorithm

SVM is mainly used to solve two kinds of problems of pattern recognition, and it regards the structure risk minimum principle and the VC dimension theory in statistical learning theory as the foundation to solve the small sample, and nonlinear and high dimensional pattern recognition problems with excellence.

The basic idea of SVM is to construct the hyper plane as a decision plane, and to makes the biggest gap between positive and negative model. After the projection of the sample point to the vector space, seek a hyper plane which can divide them. If training vector can be correctly

PoS(ISCC2015)033

divided by a planar linear and the distance between them is widest, we call it optimal hyper plane. Among them, the nearest point from the optimal hyper plane is support vector.

Assume that the training set is:

$D=(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n),x\in R^n,y\in\{+1,-1\}$, X is positive, Y is +1, conversely, Y is -1. If the data set can be divided by a linear hyper plane segmentation, the hyper plane is expressed in equation: $w\cdot x+b=0$.

According to the definition of the SVM, only when they meet the following conditions can we find out the optimal hyper plane.

$$w\cdot x\geqslant -1, y_i=+1$$

$$w\cdot x\leqslant -1, y_i=-1$$

In accordance with the above inequalities, it can be combined into the following inequality $y_i(w\cdot x+b)\geqslant 1, i=1,2,3,\ldots,l$.

To sum up, the optimal hyper plane is equivalent to solving the minimum value of $\partial(w)=\|w\|^2$, so when the classification distance equals to $2/\|w\|$, the largest interval will appear.

The VC theory of statistical learning shows that in a high-dimensional N dimensional space, assuming the sample set exists in the super ball area with radius R, a specification hyper plane subset shall meet the following inequality:

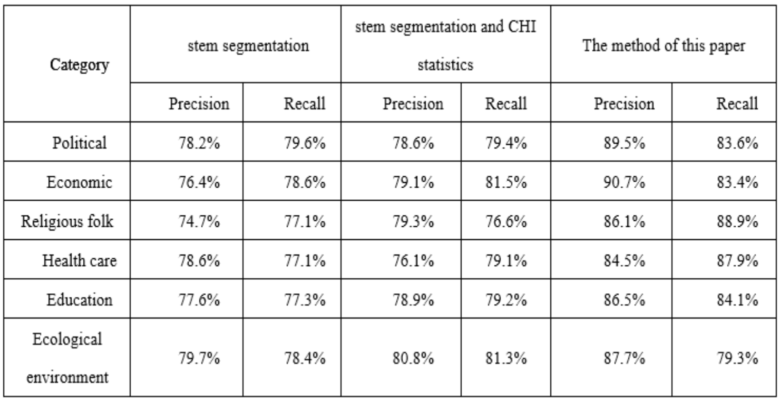
$$h\leqslant min([R^2A^2],n),\|w\|\leqslant A$$

Therefore, by minimizing $\|w\|$, reduce the degree of confidence of VC dimension to the minimum.

5. Experimental Analysis

In this paper, the corpuses are collected from nearly 3600 documents by websites in Uyghur, though the artificial way sort out the six categories, and they are: political, economic and religious customs, health, education and ecological environment. Each collection is made up of 600 for text sets and 3000 for training sets.

Classified experiment in Uyghur is mainly divided into three contrast ones, and they are: only for stem segmentation, for stem segmentation and x^2 , and for improved x^2 and model aggregation. Use the precision ratio as classified evaluation method, comparing-results are shown in the Table 2:



Category	stem segmentation		stem segmentation and CHI statistics		The method of this paper	
	Precision	Recall	Precision	Recall	Precision	Recall
Political	78.2%	79.6%	78.6%	79.4%	89.5%	83.6%
Economic	76.4%	78.6%	79.1%	81.5%	90.7%	83.4%
Religious folk	74.7%	77.1%	79.3%	76.6%	86.1%	88.9%
Health care	78.6%	77.1%	76.1%	79.1%	84.5%	87.9%
Education	77.6%	77.3%	78.9%	79.2%	86.5%	84.1%
Ecological environment	79.7%	78.4%	80.8%	81.3%	87.7%	79.3%

Table 2: Test Result of Classifier

PoS(ISCC2015)033

6. Conclusion

This article focuses on the importance of feature selection for the task of classification and proposes the method of the feature selection based on x^2 and integration of model aggregation selection. By comparing the experiment, we found that not only the classified experiment of text for feature selection method based on improved x^2 and model aggregation, has been improved significantly in the recall rate and accuracy, but also the speed of classification has increased by 25%.

As some labels in Uyghur text don't belong to the specific text category, improving the performance of text categorization should involve more factors, and the next step of work includes: (1) the expansion of Uyghur corpus; (2) the study of text categorization method based on Uyghur phrases; (3) as stem segmentation, feature extracting and other methods of research are not mature enough, it will be based on N-Gram to do the study; (4) to emerge the basic meaning of Uyghur into the procession of the text classification.

References

[1]J. S. Su, B. F. Zhang, X. Xu. *Advances in Machine Learning Based Text Categorization[J]*. journal of software.2006(9):1848-1859

[2]Gupta K M. Moore P G Aha D W. Pal S K. *Rough set feature selection-methods for case-based categorization of text documents*[C]. Proceedings of the 1st International Conference on Pattern Recognition and Machine intelligenc.20053776:792-798.

[3]Liu He, Liu Dayou. *A Feature Weighting Scheme for Text Categorization Based on Feature Importance* [J].journal of computer research and development.2009, 46(10):1693-1703.

[4]Meng Jiana, Lin Hongfei. *Based on the characteristics of the contribution of feature selection methods applied in the text classification*[J].journal of dalian university of technology. 2011, 51(4):611-615.

[5]Alimjan AYSA, Turgun IBRAHIM, Hasan OMAR. *The Uyghur text classification based on machine learning research* [J].computer engineering and application. 2012, 48(5):110-113.

[6]Mayfield J, McNamee P. *Single n-gram stemming*[C]. Efthimiadis E N, Dumais S T, Hawking D, et al. Proceedings of the 26th Annual International Retrieval. New York:ACM,2003:415-416

[7]Melucci M, Orio N. *A novel method for stemmer generation based on hidden Markov models*[C] .Kraft D, Frieder O, Hammer J, et al. Proceeding of the Twelfth International Conference on Information and Knowledge Management. New York: ACM, 2003:131-138.

[8]Batuer Aisha, Masong Sun. *A statistical Method for Uyghur Tokenization*[C]. Proc. of the 2009 IEEE International Conference on NLP-KE 2009. Dalian, 2009:383-387

[9]Aishan Wumaier, Tuergen Yibulayin, Zaokere Kadeer. Shengwei Tian *Conditional Random Fields Combined FSM Stemming Method for Uyghur Proceeding*[C]. 2nd IEEE International Confrence on Computer andinformation Technology (ICCSIT 2009). 2009.8:295-299