Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 742-747
doi:10.4028/www.scientific.net/AEF.6-7.742

Research on Building Methods of Hierarchical Structure in Text Classification

Yunbo xiong[1, a]

[1] School of Information Management, Jiangxi University of Finance and Economics, NanChang, Jiangxi, china

[a]yunbo09@gmail.com

Keywords: text classification; confusion matrix; hierarchical structure; hierarchical clustering; confusion category

Abstract. There always exists semantic hierarchical relationship in text classification. Therefore, it's inevitable to organize documents in accordance with the hierarchical structure. Based on confusion matrix, this paper attempted to adopt two different algorithms including hierarchical clustering and confusion category to build hierarchical structure of document category, and finally made use of hierarchical classification to carry on experiment, results of which showed that the confusion category strategy is superior to hierarchical clustering strategy and recall and precision of flat classification are both improved.

Introduction

In the text classification process, document categories are always regarded to be in the same plane level rather than intersectant [1]. However, in the case that a document library is particularly large with lots of document categories or document category differentia is small, or document category itself has a semantic hierarchical structure, people always tend to organize, manage and classify documents in the document library in accordance with the concept hierarchical structure. Therefore, hierarchical oriented text classification has become a hot issue in the field of text classification [1-6].

Hierarchical structure of document categorization is equivalent to a tree called as the document categorization tree. Root node of categorization tree represents the entire document library, and other nodes represent classes. Leaf node is the base class or subclass, while the other nodes are super class or parent class [1]. McCallum made use of Naive Bayes to carry on hierarchical classification, allowing documents to belong to any category that non-root node corresponds. As for the category with relatively sparse data, shrinkage techniques were used in its parameter estimation [3]. In accordance with hierarchical structure of document categorization, Koller tiered up and differentiated the hierarchical classification into every partial classification problem, and used very few features to establish classifier in each internal node of category hierarchy [4]. The premise of problems these studies are to solve is that: the document categorization to be divided has evidently semantic hierarchical relationship, and hierarchical structure is usually manually built by the user.

What the paper is to solve is: how to automatically build hierarchical structure of document categorization according to confusion matrix produced by plane classifier. The paper first introduced the basis to build hierarchical structure of document categorization: confusion matrix, and then concretely introduced two building methods of hierarchical structure of document categorization: hierarchical clustering and confusion category. Finally, according to these two methods respectively, the method of hierarchical classification was used to make test of document classification which were also analyzed and compared.

Confusion matrix

Usually, errors of statistical classifier are showed in confusion matrix. Confusion matrix is widely used in pattern recognition. According to confusion matrix, Wan Hao made use of HAC (Hierarchical Agglomerative Clustering) to build a two-tier hierarchical classifier based on error correction strategy to achieve whole aspect range HRRP recognition [5]. Based on confusion matrix of the consonants perception of Chinese synthetic speech, Zhang Jialu evaluated Chinese speech synthesis system, and with comparison of definition of natural language and synthetic language in the different language levels he studied defects of the system in terms of the prosodic features[6]. Godbole combined rapidness of Naive Bayes classifier and precision of SVM classifier, and according to the document category confusion matrix he made SVM classifier be applied to various categories classification as well as ultra-large-scale document classification [7,8].

The following is a brief introduction of confusion matrix.

A category set $C = \{c_1, c_2, \cdots, c_k\}$ is given and confusion matrix of classifier can be expressed as:

$$CM = \begin{bmatrix} n_{11} & \cdots & n_{1j} & \cdots & n_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & n_{ij} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{k1} & \cdots & n_{kj} & \cdots & n_{kk} \end{bmatrix}$$

Among them, n_{ij} refers that the number of category c_i is identified as the category c_j by classifier.. If $i = j$, n_{ij} refers that the number of samples in the category c_i is correctly identified by classifier; if $i \neq j$, n_{ij} indicates that the number of category c_i is wrongly identified as category c_i by classifier. Confusion matrix CM represents space distribution of category C, reflecting the recognition performance of the classifier: the i line responds the recall of the category c_i, while the j column reflects precision of the category c_j. Ideally, if the classifier's recall and precision are both 100%, then only the diagonal elements of confusion matrix is non-zero value.

The confusion matrix of classifier can be gotten if training data is used to build a classifier which is then tested in the training document set. For example, Table 1 is confusion matrix by testing on 934 documents in 10 categories using the KNN method. Of course the above confusion matrix can also be normalized.

Table 1 An example of confusion matrix

Category Name		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Environment	**1**	51	0	1	0	5	3	1	1	1	4
Computer	**2**	1	56	1	0	7	0	0	0	0	1
Transportation	**3**	0	0	61	0	4	3	0	0	0	3
Education	**4**	0	0	0	65	3	0	1	0	2	2
Economy	**5**	0	0	0	0	101	0	2	0	0	5
Military	**6**	1	0	0	2	1	51	4	0	0	24
Sports	**7**	1	0	0	0	0	0	148	0	0	0
Medicine	**8**	6	0	0	1	4	0	1	53	1	2
Art	**9**	0	0	0	3	1	1	0	0	75	2
Politics	**10**	0	0	0	0	4	1	1	0	1	160

Hierarchical structure

A document hierarchical structure can be built by making confusion matrix as a priori knowledge, that is, the hierarchical structure category tree. This paper presents two kinds of hierarchical structure building methods.

Using hierarchical clustering algorithm to build hierarchical structure. Usually there are two kinds of hierarchical clustering algorithms: agglomerative and divisive. The agglomerative algorithm starts from n subcategory which each contains only one data point. In each step of the algorithm, agglomerate the two most similar categories to form a new cluster, thus the number of clusters every time reduces one, and when all the data fall into a cluster, the algorithm ends. Divisive algorithm starts in a major category, and gradually be divided until the formation of n categories, with each category a separate individual. Usually, divisive algorithm has low computational efficiency [1]. This paper uses agglomerative algorithm. The steps are as follows:

(1) in confusion matrix, a row or column each plane category corresponds to forms a vector as individual of the category in the clustering; n categories form n individuals (as n category) and then calculate the mutual distance d_{pq} among n individuals, thus forming a $n \times n$ symmetric matrix $D_{(0)}$ with its diagonal elements 0, as $d_{pq} = d_{qp}$.

(2) Select the smallest element in $D_{(0)}$ in addition to the diagonal elements (set to be d_{pq}), and agglomerate the categories C_p and C_q into a new category C_r. Eliminate the corresponding rows and columns of C_p and C_q in $D_{(0)}$, and add a row and a column composed by the new category C_r and other non aggregation categories to get a new distance matrix $D_{(1)}$. Clearly, $D_{(1)}$ is a $(n-1)\times(n-1)$ matrix.

(3) start from $D_{(1)}$ and repeat step (2) to get $D_{(2)}$, and so on, until the n categories are clustered into a major category.

Fig.1 is category hierarchical structure after clustering data of the confusion matrix in Table 1 using SPSS.

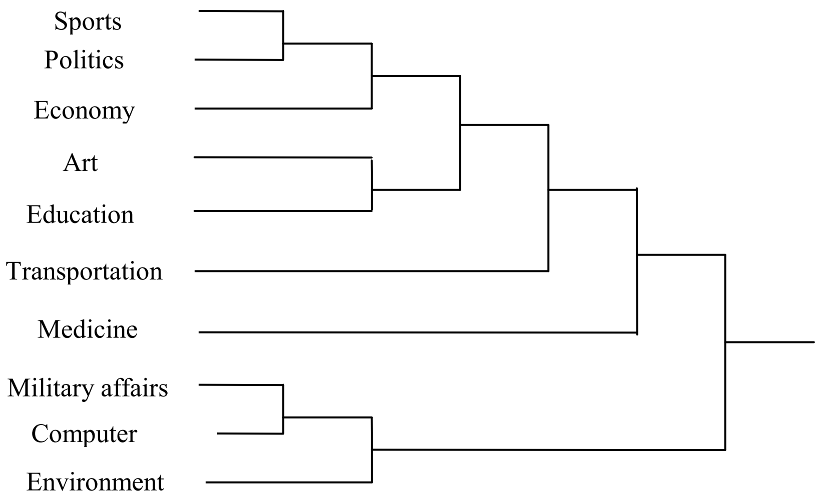

Fig.1 Building categorization tree by clustering

Using confusion category algorithm to build hierarchical structure of document categorization

Definition 1: Given a plane category set $C = \{c_1, c_2, \cdots, c_k\}$ and a confusion matrix CM, n_{ij} represents the element in i row, j column in the confusion matrix, and the probability that samples in category c_i are identified as category c_j,which is expressed as:

$$P_1(c_i \mid c_j) = n_{ij} / \sum_{j=1}^{k} n_{ij} \tag{1}$$

And in the condition that the recognition result is c_i, the probability of samples from the category c_j is:

$$P_2(c_i \mid c_j) = n_{ij} / \sum_{i=1}^{k} n_{ij} \tag{2}$$

Definition 2: Given the threshold value $t(0 \le t \le 1)$, and define the confusion category $CF_1(c_i)$ and $CF_2(c_i)$ of category *i* respectively as all the categories whose identification probability $P_1(c_i \mid c_j)$ and $P_2(c_i \mid c_j)$ is larger than t, which is expressed as:

$$\text{If } P_n(c_i \mid c_j) \ge t \text{, then } c_j \in CF_n(c_i) \text{ (n = 1,2)} \quad (3)$$

According to the above definitions, if category B is the confusion category of category A, category A is not necessarily that of category B, that is, confusion category does not have symmetry.

$CF_1(c_i)$ considers recall of the category c_i, while $CF_2(c_i)$ considers the precision. These two confusion categories can be used to build the category hierarchical structure. However, if a different confusion category is used, the category hierarchical structure is also different, because they are considered from different angles. For example, set threshold value t to be 5%, according to the confusion matrix in Table 1, confusion category of the "environment" category is CF_1 (environment) = {economy, politics} and CF_2 (environment) = {medicine}. It can be seen that the confusion category acquired by using $CF_2(c_i)$ is more in line with the intuitive image of people, and the experimental results also proved that the category hierarchical structure built by using $CF_1(c_i)$ would reduce the classifier precision, as $CF_1(c_i)$ pays more attention to the recall. Therefore, the paper used the confusion category $CF_2(c_i)$ which better shows precision as confusion category to build document category hierarchical structure.

According to confusion category in each category, fix on threshold value t, then a two-tier category hierarchical structure can be built.

The first layer: composed by *k* nodes, respectively represent *k* plane categories in the plane category set;

The second layer: as for category c_i each node of the first layer corresponds to, if $CF_2(c_i)$ is empty, then this node does not exist a sub node, otherwise, its sub node is category $i \cup CF_2(c_i)$.

The selection of the threshold value t is critical. If t is too large, confusion category in many categories is the empty set. If t is too small, confusion category in many categories will be so many to increase the number of hierarchical classifier, affecting the speed of classification. It was found that t is generally more appropriate to select from 3% to 5%. Fig.2 is the category hierarchical structure built for confusion matrix in Table 1, when t takes the value 4%.

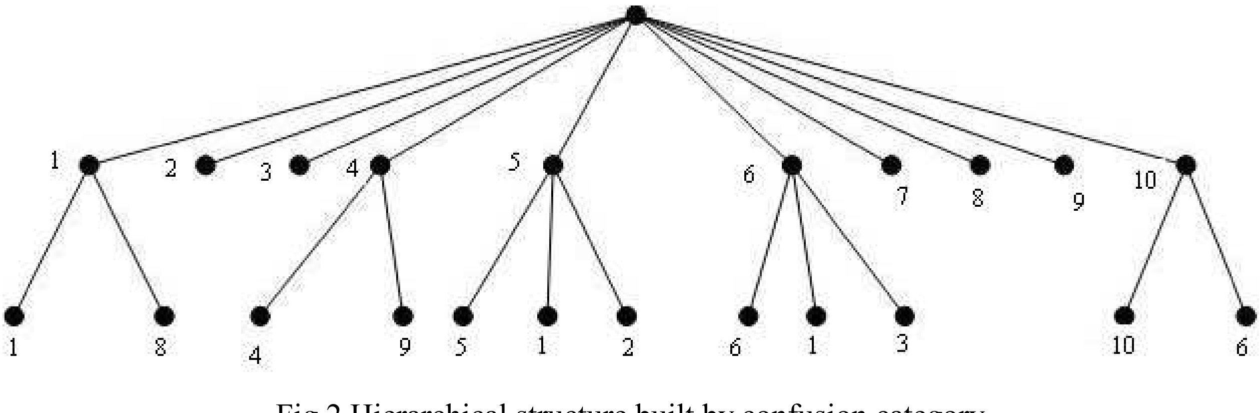

Fig.2 Hierarchical structure built by confusion category

Hierarchical classification

There is no difference between hierarchical classification and plane classification; first select the classifier, and then make classifier training. However, hierarchical classification still has its unique characteristics.

In the learning stage of hierarchical classification, build a classifier respectively for each internal node in the document category hierarchy tree; in the classification stage, upon the arrival of a document to be classified, firstly assign it to some category with the top-layer classifier, then use the classifier to continue to classify documents, and proceed it until the documents are assigned to a base class. The problem of this classification process is that: if the upper-layer classifier made a wrong classification, mistake would continue to the end. A way to solve this problem is to consider taking a number of classification paths when classifying downwards, and then to compare results through different paths. A safer method is to consider all of the classification paths, and finally choose an optimal classification path, then the bass class the classification corresponds to is the category where the document is. The specific hierarchical classification method is showed in References [1].

Experimental results

The document sets in the experiment were divided into two parts: the document set D1 and document set D2. D1 is composed of 2815 documents in 10 categories, including 1881 training documents and 934 test documents, with training documents more evenly distributed; D2 is composed of 3395 documents in 38 categories, including 2256 training documents and 1139 test documents with training documents unevenly distributed, that is, some are less than 20, while some as many as 200.

Fig.3 shows experimental results when t varies by building hierarchical structure based on confusion category aimed at the document set D1 with hierarchical classification.

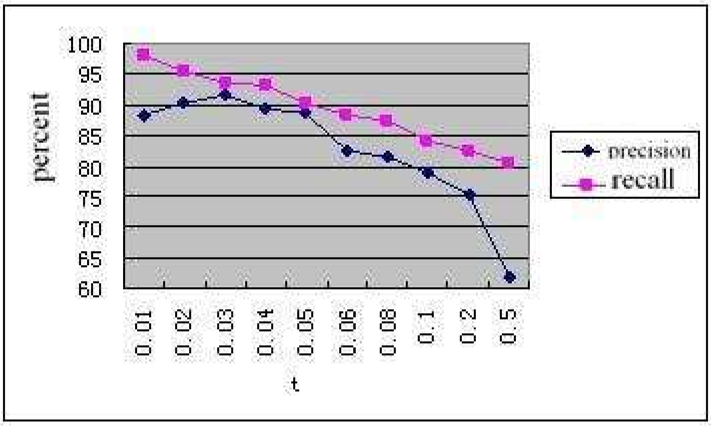

Fig.3 Classification performance when t varies

It can be seen from Fig.3 that there is a clear performance change for hierarchical classifier when t varies. Generally, with the gradual increase of t, the precision and recall of classifier gradually decrease with the recall rate reducing relatively flat, while the precision decreases more significantly. In addition, it is worth noting that: when t is in the interval [0.01,0.03] precision increases as t increases. This is because that when t is too small, the following two questions will occur:

(1)confusion matrix is produced by some plane classifier, during the classification which may wrongly assign some documents to some category due to inappropriate parameters selection of the plane classifier or rather special certain documents, while during hierarchical classification, the classifier may be able to correctly identify these documents, so if t is too small, it will actually affect precision and recall rate of the hierarchical classification.

(2) if t is too small, confusion category of many categories will be numerous, which on one hand has increased the complexity of hierarchical classifier and affected the speed of hierarchical classification and on the other hand, also has increased the error probability of hierarchical classifier.

Therefore, in the experiment, the values from 0.03 to 0.05 as for *t* are more appropriate.

In addition, in order to evaluate and compare the above two hierarchical structure building methods, the paper made related experiments. In the experiment, the hierarchical classification uses two methods of KNN and SVM, and performance evaluation standard adopts precision and recall.

The results are shown in Table 2 where there lists the average precision of classifier in terms of document sets D1 and D2 with the use of different classification methods and different classification strategies.

Table 2 Performance Comparison for different document-sets and classification algorithms

Document set	KNN			SVM		
	Plane classification	Hierarchical classification		Plane classification	Hierarchical classification	
		Confusion category	Hierarchical clustering		Confusion category	Hierarchical clustering
D1	87.9	89.2	85.5	95.3	95.8	93.1
D2	71.5	72.7	66.8	83.1	84.2	79.8

Conclusions and summary

This paper described confusion matrix of the classifier, and built hierarchical structure of document category by using hierarchical clustering and confusion category which are analyzed and compared based on confusion matrix. Experimental results showed that: (1) hierarchical structure of document category built by confusion category method is significantly superior to that of hierarchical clustering method, mainly because the confusion category method takes account of asymmetry among the confusion categories; (2) using the same classification methods(like KNN), result of confusion category strategy is better than that of plane classification, while the classification performance of hierarchical clustering strategy is related to the specific application.

References

[1] Yuan Shijin, Li Ronglu, Zhou Shuigeng, Hu Yunfa. Hierarachical Chinese Document Categorization.Journal of China Institute of Communications,Vol.25 No.11,2004:55-63

[2] Zhan Xuegang,Lin Hongfei,Yao Tianshun.Hierarachical Method for Chinese Document Classification.Journal of Chinese Information Processing,Vol.13,No.6,1999:20-25

[3] McCallum.A,Rosenfeld.R,Mitchell.T,Ng.A.Improving text classification by shrinkage in a hierarchy of classes.In:Proceedings of the 15th International Conference on Machine Learning (ICML98).Morgan Kaufmann Publishers Inc,San Francisco,CA,USA,1998,359-367

[4] Koller.D,Sahami.M.Hierarchically classifying documents using very few words.In:Proceedings of the 14th International Conference on Machine Learning (ICML97),Morgan Kaufmann Publishers Inc.San Francisco,CA,USA,1997,170-178

[5] Wan Hao,Ren Yong,Shan Xiuming.Confusion-Matrix Based Whole Aspect Range HRRP Recognition,Microelectronics & Computer,Vol.22,No.3,2005:136-143

[6] Zhang Jialu,Qi Shiqian,Yu ge.Assessment methods of speech synthesis systems for Chinese,Acta Acustica,Vol.23,No.1,1998: 19-30

[7] Godbole.S.Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers.Technical Report,Indian Institute of Technology,Bombay, 2002,Available online at http://citeseer.nj.nec.com/godbole02exploiting.html

[8] Godbole,Sarawagi.S,Chakrabarti.S.Scaling multi-class support vector machines using inter-class confusion.In:Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.ACM Press,New York,NY,USA,2002,513-518.