Oriented to cloud service environment Multi dimension query of English Network

LIU Jing[1], MA Zhenyu[2]

1. School of Fundamental Studies, Shanghai University of Engineering Science, Shanghai 201620, China

2. China Europe International Business School, Shanghai, 201206, China

Email:jingmonkey@126.com

Keywords: multimedia resource; video resource retrieval; SVM; English teaching; video

Abstract. The arrival of the information age makes people more and more exposed to the multimedia information, and the multimedia teaching based on the visual audio and graphic images has been widely used. On English multimedia teaching information in text data content analysis and retrieval technology is relatively mature, such as Internet search engines is the keyword based retrieval approach, since the content data with structural features, so it can be used to describe the relationship model. And video, audio and other multimedia information content with the characteristics of unstructured, is not easy to describe with relational model, coupled with the video and audio is a relationship with the time of continuous media information, network in the video and audio flow media in the form of existence. Therefore, the flow media form of retrieval is very difficult. In this paper, we propose a video resource retrieval method for English teaching.

Introduction

With the rapid development of network technology and exploding increase of various network information, more and more education-related English teaching video resources into the Internet, these network education English teaching video resources information become an important source of people’s access to information increasingly. How to quickly, effectively and quickly obtain the required basic education English teaching video resources, distinguish them with other English teaching video resources, and classify the information are the main research content of this paper. eb text classification automatically determines the type of text based on text content and attributes. Large amounts of text are subjected to a subject or multiple categories, this paper analyzes the classical SVM (support vector machine) algorithm, and finally puts forward the weighted multi-class SVM[3] algorithm, by comparing with the experimental results, it shows that this algorithm has a good effect in the classification of basic education English teaching video resources.

Support vector machine SVM

Support vector machine (SVM) is a machine learning algorithm based on the principle of minimum structural risk and statistical learning theory, can effectively solve the pattern recognition problem under high dimension, nonlinear and local small sample, this algorithm has been widely used in face recognition [5], fingerprint recognition, text classification, and other fields and achieved good effects. However the SVM algorithm was originally designed for solving two kinds of problem, when dealing with multiple-class problem, it needs to construct a suitable classifier. At present, the structure of the classifier mainly implemented by combining multiple classifiers, SVM includes two important concepts, one is the optimal interval classifier and the other is kernel function. The existing multi-class SVM methods [2] are: one-against-one SVM (one-to-one) and one-against-rest SVM (one-to-many), binary tree support vector machine (BT- SVM), directed acyclic graph support vector machine (DAG - SVM).This paper mainly conducts improvement to SVM algorithm based on the one-to-rest method.

One-to-rest constructs decision-making between class A sample and the rest multi-class sample, in training, in turn, to classify a certain category of sample as a class, so k categories of sample can construct k SVMs. When constructing the SVM classifier for class I, take training data of I as positive vector, the rest of the training data as a negative vector, and construct a decision-making boundary for class I and the rest of the classes, through two kinds of SVM to determine a decision function, so a total of k decision functions. Given a test sample x, to determine the value of k decision functions respectively, if the value of ki is maximum, then x belongs to class i. The number of decision boundaries constructed through one-to-rest method is less, so its prediction speed is faster than that of one-on-one SVM algorithm. But because it needs to calculate all sample set when constructing decision boundary each time, so the time spent on training is much more.

Classification algorithm of educational English teaching video resources

Before the analysis of the weighted multi-class SVM algorithm, the theory and kernel function of two kinds of SVM shall be understood first. In solving the problem of classification, although the SVM is a kind of effective method, but there are still some defects, because it needs to train all data, while in many practical applications, not all of the data are critical for classification, because the general collected data contains a lot of noises and outliers. And SVM is sensitive to the noise data and outliers, training is focused on some points may be far away from the actual position and even distributed on the error side of the feature space. In the process of training, singular point with larger Lagrange multiplier can be converted into support vector. So many improved Support Vector Machine (SVM) such as RSVM [6] (Robust Support Vector Machine), SVND [7] (Fuzzy Support Vector those), FSVM [8] (Fuzzy Support Vector those) are used to solve this problem. The main idea of the SVM model is by translating the original optimization problem into a quadratic programming problem to construct the optimal decision boundary. Training data set, such as equation(1), where （xi,yi）,xi∈RN.

$$T=\left(x_1,y_1\right),\left(x_2,y_2\right),\ldots,\left(x_N,y_N\right)\in R^N\times\left\{+1,-1\right\} \qquad （1）$$

Original optimization problem:

$$\min\left(w,\xi,\right)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \qquad （2）$$

$$y_i\left(\left(w.x_i\right)+b\right)\geq 1-\xi_i, i=1,2,\ldots N, \qquad （3）$$

$$\xi_i\geq 0, i=1,2,\ldots N \qquad （4）$$

In above equation（2）, $\xi=(\xi_1,\ldots\xi_N)^T$ ，C>0，is a penalty parameter. Not only to minimize‖w‖2, but also minimize $\sum_{i=1}^{N}\xi_i$.

Weighted SVM algorithm makes a difference among these data according to different weights of each data. Weighted SVM gives higher weight to the data with important information, similarly, the data carried less important information is given less weight. Weighted training data set:

$$T=\left(x_1,y_1,v_1\right),\left(x_2,y_2,v_2\right),\ldots,\left(x_N,y_N,v_N\right)\in R^N\times\left\{+1,-1\right\}$$

Where, $\varepsilon<v_i<1$ is the weight of（zi, yi）(i=1,2,...N), small enough positive number of ε ， xi∈RN. The same as SVM，weighted SVM mainly by maximizing classification interval, minimizing classification error rate to achieve classification accuracy. Different from SVM is that the weighted SVM using a weighting function to weaken some unimportant data to enhance the influence of important data. Under the condition of data weighting to construct optimal decision boundary, the optimization problem is converted into：

$$\min\left(w,\xi,v\right)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}v_i\xi_i \qquad （5）$$

$$y_i\left(\left(w.x_i\right)+b\right)\geq 1-\xi_i, i=1,2,\ldots N, \qquad （6）$$

$$\xi_i \geq 0, i = 1, 2, \ldots N \qquad (7)$$

It can be seen from the above equation（5）, introduced ν I decreases $\sum_{i=1}^{N} \xi_i$ to a large extent, decreases the influence of slack variable ξ_i in optimization problem, thus the（xi, yi）can be seen as the less important data for classification.

The above weighted optimization problem can be converted into a convex quadratic programming problem:

$$\max \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j \left(x_i . x_j \right) \qquad (8)$$

$$s.t. \qquad \sum_{i=1}^{N} a_i y_i = 0, \qquad (9)$$

$$0 \leq a_i \leq v_i C, i = 1, 2 \ldots . N \qquad (10)$$

K（ .）=$\left(x_i . x_j \right)$ is kernel function, this paper adopts radial basis function（RBF） with better learning capacity and wider convergence domain.

It can be seen that, assuming ν i =1,WSVM is converted into original problem of support vector machine, for different ν I can determine the compromise of xi in the system. The smaller of ν i, and the importance of xi for constructing maximal margin hyperplane is less, v.v.

Video classification query

In order to obtain basic education English teaching video resources information from the network, one shall use web crawler tools to grab some web pages at first, then put these text stored in the local resource system. But most of the web papers are in the form of hypertext HTML, not only contain the thematic information and also contain a large number of symbols and links, so in order to effectively extract the feature information, one must proceed preprocessing to a web page. Web pretreatment is mainly to extract the title and content of basic education English teaching video resources in order to get text-only file. However, the obtained text contains a lot of function words, stop words. So one needs to use tokenizer to get information related to basic education English teaching video resources. Due to the large amount of text, text vector space is also big, in order to solve the problem of data sparseness, this paper combines document frequency and mutual information for feature extraction of the text. In the process of feature selection, to choose representative features to represent the document information, and reduce the feature dimension. In order to let the machine can carry on the processing and calculation to the text, this paper uses the vector space model to construct the model of document classification. Finally, it adopts TF-IDF to calculate weight, and establish sample space as the training set.

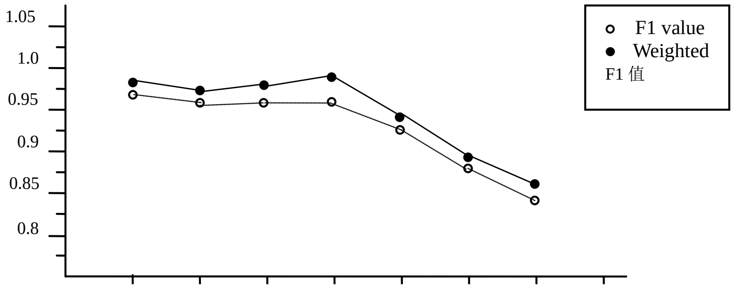
Experiment effect and analysis

Downloaded 3142 video from the Internet, and are divided into seven classes, among them, 2411 video as the training document set, remains as the test set, as shown in table 1.This experiment adopts the commonly used evaluation method, recall rate, precision rate, and F1 value. Recall rate refers to ratio of the number of retrieved documents and the total number of relevant documents, precision rate is the ratio of the number of retrieved documents in system and the total number of returned documents in system. F1 value is the most commonly used method to measure the effect of the overall classification.

$$F_1 = \frac{\Pr ecision \times \operatorname{Re} call \times 2}{\Pr ecision + \operatorname{Re} call} \qquad (11)$$

Table 1 training sample and test sample

Class	Primary dialog class	Intermediate dialog class	Advanced dialog class	Interpretation of ball games	Voice of America	American spoken	American slang
Training sample	450	447	276	342	296	292	308
test sample	143	144	100	135	88	60	61



Primary dialog class, Intermediate dialog class, Advanced dialog class, Interpretation of ball games

Conclusion

This paper proposes a weighted multi-class SVM algorithm and its application in the classification of education English teaching video resources. The basic idea of weighted SVM training designs a weighted multi-class SVM model, which can really affect the noise distribution in data set. Data with important information have larger weights, noise data and outliers have lower weights, so the weighted SVM constructs decision boundary according to the importance of training data in training set. And the experiment of education English teaching video resources classification demonstrates the weighted SVM algorithm is obviously superior to the multi-class SVM classification algorithm.

Reference

[1] Jinyu Hu, Zhiwei Gao and Weisen Pan. Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation[J]. Journal of Applied Mathematics, 2013 (2013).

[2] Lv Z, Tek A, Da Silva F, et al. Game on, science-how video game technology may help biologists tackle visualization challenges[J]. PloS one, 2013, 8(3): e57990.

[3] Jiang D, Ying X, Han Y, et al. Collaborative multi-hop routing in cognitive wireless networks[J]. Wireless Personal Communications, 2016, 86(2): 901-923.

[4] Lin Y, Yang J, Lv Z, et al. A self-assessment stereo capture model applicable to the internet of things[J]. Sensors, 2015, 15(8): 20925-20944.

[5] Jinyu Hu and Zhiwei Gao. Distinction immune genes of hepatitis-induced heptatocellular carcinoma[J]. Bioinformatics, 2012, 28(24): 3191-3194.