International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.73-78
http://dx.doi.org/10.14257/ijdta.2016.9.8.08

Application of Weighted Multi-Feature Selection in Educational Resources Classification

Wen Ying[1] and Li Hao[2]

[1]*Changsha University of Science & Technology, Huanan Changsha, 410015*
[2]*Hunan zhongyi communication technology Engineering Co., Ltd., Huanan Changsha, 410015*
395208367@qq.com

Abstract

Support vector machine (SVM) has been widely applied to small-sample, non-linear and high-dimensional classifications. Many modified SVM algorithms were put forward in recent years. Some of them focus on SVM feature selection and some focus on SVM classification effectiveness. As different input vectors have significant influence on learning effect of decision boundary, this paper proposes a weighted multi-class support vector machine (WSVM) algorithm. The algorithm gives different weights to features according to the importance of their information. WSVM algorithm establishes decision boundaries based on weights and is used to classify educational resources. Experimental results indicate that the method achieves relatively good classification effectiveness.

Keywords*: Educational resources; Support vector machine (SVM); Weight; Text classification*

1. Introduction

More information about educational resources are connected to the Internet with rapid development of network technologies and explosion of network information. These education resources have been major sources for people to gain information. This paper focuses on distinguishing basic educations resources from other resources and information classification. Web texts automatically fall into different categories according to their contents and properties. Plenty of texts is subject to one theme or several categories. The paper analyzes the classic SVM algorithm and proposes weighted multi-class support vector machine (WSVM) algorithm which is applied to multi-class classification. Experimental results indicate that the method achieves relatively good effectiveness in classification of basic educational resources.

2. Support Vector Machine (SVM)

SVM is a machine learning algorithm based on the theory of minimum structuring risk and the statistical learning theory. It is an effective solution to high-dimensional, non-linear and small-sample pattern recognition problems and has been widely used in fields such as face recognition, fingerprint recognition and text classification with good application results. As two-class SVM algorithm is designed to solve two-class problems, a suitable classifier needs to be structured to deal with multi-class problems. At present, the main method to structure is combining several classifiers. SVM includes two important concepts: optimal margin classifier and kernel function. Existing multi-class SVM algorithms include one-against-one SVM, one-against-rest SVM binary tree SVM (BT-SVM) and directed acyclic graph SVM (DAG-SVM). This paper modifies the one-against-rest

ISSN: 2005-4270 IJDTA

SVM which structures decision boundaries between one-class samples and multi-class samples. Samples belonging to a category are classified into same SVM, so k SVMs are structured for k categories of samples. When structuring SVM classifier for i class, i training data is taken as positive vector and the rest training data are taken as negative vectors. Then a decision boundary is established for i category and two-class SVM is used to drive a decision function. Thus there are k decision functions in total. Specify a test sample x and calculate the values of k decision functions. If ki has the maximum value, x belongs to i category. The number of decision boundaries structured by one-against-rest SVM is small, so the prediction speed is faster than one-against-one SVM. But all sample sets need to be calculated when it is used to structure decision boundary, the training process takes more time.

3. Weighted Multi-Class Support Vector Machine (WSVM)

It is necessary to know the theory and kernel function of two-class SVM before analyzing WSVM. Although SVM is an effective way to do classification, it is flawed as all data need to be trained. But not all data are important to classification in many fields because the collected data generally contain a plenty of noise and abnormal values. SVM is sensitive to noise data and abnormal values. Some training points may be far away from their genuine locations or even at wrong side of characteristic space. During the process of training, singular points with large Lagrangian multipliers may be transformed into support vectors. Thus, many modified SVMs such as RSVM (Robust Support Vector Machine), SVND (Fuzzy support vector machines) and FSVM (Fuzzy support vector machines) are used to solve these problems. Main theory of SVM model is transforming optimization problems to quadratic programming problems and structuring decision boundary. Training dataset is shown as in formula (1), where (xi,ci),xi∈RN.

$$T=(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\in R^N\times\{+1,-1\} \tag{1}$$

Original optimization problems are as shown in formula (2), (3) and (4):

$$\begin{cases} \min(w,\xi,)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i & (2)\\ y_i\left((w.x_i)+b\right)\geq 1-\xi_i, i=1,2,...N, & (3)\\ \xi_i\geq 0, i=1,2,...N & (4)\end{cases}$$

In the formula (2), $\xi=(\xi_1,...\xi_N)^T$, C>0 is a penalty parameter. It is necessary to minimize ‖w‖2 and $\sum_{i=1}^{N}\xi_i$.

Weighted SVM treats data according to their weights. The algorithm gives high weight to data containing important information and low weight to data containing less important information. Weighted training dataset is:

$$T=(x_1,y_1,v_1),(x_2,y_2,v_2),...,(x_N,y_N,v_N)\in R^N\times\{+1,-1\} \tag{5}$$

Where $\varepsilon<v_i<1$ is the weight of (xi, ci)(i=1,2,...N) and ε is a positive number that is small enough, xi∈RN. Like SVM, weighted SVM also achieves classification accuracy mainly through maximization of classification intervals and minimization of classification errors. Unlike SVM, weighted SVM introduces a weighting function to reduce the weight of less important data and increase the influence of important data. Weighted data are used to establish optimal decision boundary. Optimization problems are transformed into formula (6), (7) and (8):

$$\begin{cases} \min(w,\xi,v)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}v_i\xi_i & (6)\\ y_i\left((w.x_i)+b\right)\geq 1-\xi_i, i=1,2,...N, & \\ \xi_i\geq 0, i=1,2,...N & \end{cases}$$

(7)

(8)

As shown in the formula (6), introduction of ν i decreases $\sum_{i=1}^{N}\xi_i$ to a great extent and the influence of slack variable ξ_i on optimization. Thus (xi, ci) can be regarded as less important data to classification.

Weighted optimization above can be transformed to convex quadratic programming, as shown in formula (9), (10) and (11):

$$\begin{cases} \max \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j \left(x_i . x_j\right) & (9) \\ s.t. \quad \sum_{i=1}^{N} a_i y_i = 0, & (10) \\ 0 \le a_i \le \nu_i C, i = 1,2....N & (11) \end{cases}$$

K(.)=$\left(x_i . x_j\right)$is a kernel function, where . is inner product. This paper employs radial basis function (RBF) that has strong learning ability and wide convergence domain.

WSVM is transformed to original SVM problem when assuming . For different ν i, the trade-off of xi in the system can be determined. Smaller ν i indicates less importance of xi to structuring maximal margin hyperplane, and vice versa.

4. Key Technologies to Text Classification

Web crawlers are used to capture webpages in order to collect information about basic educational resources, then texts are saved in local resource system. However, most webpages are in the form of HTML. They contain both thematic information and symbols and links. Thus webpages need to be preprocessed in order to effectively extract characteristic information. Preprocessing of webpages is to extract titles and contents about basic educational resources and then to get text-only files. But texts collected from webpages contain a lot of irrelevant function words and stop words. Tokenizers are used to recognize words about basic educational resources. The paper employs document frequency and mutual information to extract characteristic texts in order to solve data sparseness caused by large text quantity and vector space. Representative characteristics are chosen to represent text information and reduce characteristic dimension. To enable the machine to process and calculate texts, the paper employs vector space model to specify the model for document classification. TF-IDF is used to calculate weights and establish sample space as training dataset.

Weighted SVM produces weights by probabilistic method. Multi-class classification problems involving two-class SVM and decision strategy are solved. Weighed SVM in this paper uses one-against-rest strategy which requires a small space. It can be described as below: containing N sample training datasets in M categories belonging to ci(xi,ci). ci (i=1,2,3,…,M) denotes the class label of sample xi. Based on one-against-rest SVM algorithm, weighted SVM is established for category k according to training sample and their expected output (xi, ci). The expected output of training sample xi is defined as below:

$$y_i = \begin{cases} +1 \quad if \quad c_i = k \\ -1 \quad if \quad c_i \ne k \end{cases} \qquad (12)$$

Output yi=+1 is called positive sample and yi=-1 is called negative sample. If Nk is the sample number of K category, the weighted SVM for K category is defined as below: grouping positive samples into H1 and negative samples into H2, then:

$$p(H_1) = \frac{N_k}{N} \tag{13}$$

$$p(H_2) = \frac{N - N_k}{N} \tag{14}$$

Where P(dj)j=1,2 represents the prior probability of the sample belonging to dj. Weight of positive training sample xi is calculated as below:

$$v \equiv p(H_i \mid x_i) = \frac{P(x_i \mid H_i)P(H_i)}{\sum_{i=1}^{2} P(x_i \mid H_i)P(H_i)} \tag{15}$$

Where P(Hj|xi)is called posterior probability and P(xi|Hj)is contingent probability meeting $\sum_{i=1}^{2} P(H_i \mid x_i) = 1$. This paper uses posterior probability as weight to train weighted SVM, then Gaussian probability-density function is used to calculate contingent probability and posterior probability.

5. Experiment Effect and Analysis

3,142 articles are downloaded from the Internet and classified into 7 categories. 2,411 articles are used as training document sets and the rest as test sets, as shown in Table 1. This research uses common evaluation methods such as recall precision and F1 value, as shown in formula (16). Recall refers to ratio of the number of retrieval documents to the total number of all documents. Precision refers to ratio of the number of retrieval documents to the number of returned documents. F1 value is the most commonly used method to evaluate overall classification effectiveness.

$$F_1 = \frac{\mathrm{Pr}ecision \times \mathrm{Re}call \times 2}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{16}$$

Table 1. Training Samples and Test Samples

Category	Chinese	Math	Physics	Chemistry	Geography	History	Politics
Training samples	450	447	276	342	296	292	308
Test samples	143	144	100	135	88	60	61

(1) Results of multi-class SVM classification, as shown in Table 2

Table 2. Results of Multi-Class SVM Classification

Category	Chinese	Math	Physics	Chemistry	Geography	History	Politics
Recall	0.950	0.936	0.969	0.976	0.897	0.924	0.942
Precision	0.993	1.000	0.969	0.964	0.971	0.853	0.796
F1 value	0.971	0.967	0.969	0.970	0.933	0.887	0.863

(2) Results of weighted multi-class SVM classification, as shown in Table 3.

Table 3. Results of Weighted Multi-Class SVM Classification

Category	Chinese	Math	Physics	Chemistry	Geography	History	Politics
Recall	0.962	0.950	0.980	0.992	0.908	0.926	0.948
Precision	0.996	1.000	0.976	0.995	0.984	0.867	0.806
F1 value	0.978	0.974	0.978	0.993	0.944	0.896	0.871

(3) F1 comparison between two algorithms, as shown in Figure 1

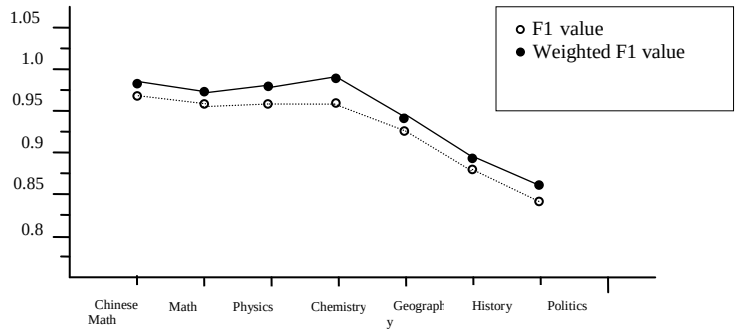

Figure 1. F1 Comparison

(4) Precision comparison, as shown in Table 4

Table 4. Precision Comparison between Two Algorithms

Algorithm	Precision
Weighted multi-class SVM	96.64%
Multi-class SVM	92.79%

6. Conclusions

This paper introduces a weighted multi-class SVM algorithm to classification of educational resources. The basic design idea is from training of weighted SVM. It has actual influence on noise distribution in datasets. Large weight is given to data containing important information while small weight is given to noise data and abnormal values. Thus weighted SVM structures decision boundary according to importance and training data. As shown in Table 2, 3 and 4 and Figure 1, weighted multi-class SVM has significant advantages over multi-class SVM.

Acknowledgement

The research is supported by Hunan education science planning fund project (XJK015QTW003) and Hunan Provincial Department of Education College Teaching Reform Project: MOOC concept of visual communication design professional practice curriculum design research.

References

[1] G. Bao, L. Mi, Y. Geng and K. Pahlavan, "A computer vision based speed estimation technique for localizing the wireless capsule endoscope inside small intestine", 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), **(2014)**.
[2] X. Song and Y. Geng, "Distributed community detection optimization algorithm for complex networks", Journal of Networks, vol. 9, no. 10, **(2014)**, pp. 2758-2765.
[3] M. Zhou, G. Bao, Y. Geng, B. Alkandari and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy", 2014 7th International Conference on Biomedical Engineering and Informatics (BMEI), **(2014)**.
[4] T. Su, Z. Lv and S. Gao, "3D seabed: 3d modeling and visualization platform for the seabed[C]. Multimedia and Expo Workshops (ICMEW)", 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 1-6.
[5] Y. Geng, J. Chen, R. Fu, G. Bao and K. Pahlavan, "Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine", IEEE transactions on mobile computing, vol. 1, no. 1, **(2015)**, pp. 1-15.
[6] Z. Lv, A. Halawani and S. Feng, "Multimodal hand and foot gesture interaction for handheld devices", ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 1, **(2014)**, pp. 10.
[7] T. Su, W. Wang and Z. Lv, "Rapid Delaunay triangulation for randomly distributed point cloud data using adaptive Hilbert curve", Computers & Graphics, vol. 54, **(2016)**, pp. 65-74.
[8] J. Hu, Z. Gao and W. Pan, "Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation", Journal of Applied Mathematics, vol. 2013, **(2013)**.
[9] S. Zhou, L. Mi, H. Chen and Y. Geng, "Building detection in Digital surface model", 2013 IEEE International Conference on Imaging Systems and Techniques (IST), **(2012)**.
[10] J. He, Y. Geng and K. Pahlavan, "Toward Accurate Human Tracking: Modeling Time-of-Arrival for Wireless Wearable Sensors in Multipath Environment", IEEE Sensor Journal, vol. 14, no. 11, **(2014)**, pp. 3996-4006.
[11] Z. Lv, A. Halawani and S. Fen, "Touch-less Interactive Augmented Reality Game on Vision Based Wearable Device", Personal and Ubiquitous Computing, vol. 19, no. 3, **(2015)**, pp. 551-567.
[12] G. Bao, L. Mi, Y. Geng, M. Zhou and K. Pahlavan, "A video-based speed estimation technique for localizing the wireless capsule endoscope inside gastrointestinal tract", 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), **(2014)**.
[13] D. Zeng and Y. Geng, "Content distribution mechanism in mobile P2P network", Journal of Networks, vol. 9, no. 5, **(2014)**, pp. 1229-1236.
[14] W. Gu, Z. Lv and M. Hao, "Change detection method for remote sensing images based on an improved Markov random field", Multimedia Tools and Applications, **(2015)**, pp. 1-16.
[15] Z. Chen, W. Huang and Z. Lv, "Towards a face recognition method based on uncorrelated discriminant sparse preserving projection", Multimedia Tools and Applications, **(2015)**, pp. 1-15.
[16] J. Hu and Z. Gao, "Distinction immune genes of hepatitis-induced heptatocellular carcinoma", Bioinformatics, vol. 28, no. 24, **(2012)**, pp. 3191-3194.

Authors

Wen Ying, received her M.S. degree in agricultural extension from Central South University of Forestry and Technology in Changsha, China. She is currently a lecturer in the College of Art & Design at Changsha university of Science &Technology. Her research interest is mainly in the area of Educational model about Mooc, Interaction design. She has published several research papers in scholarly journals in the above research areas and has participated in several books.

