Advances in Intelligent Systems Research, volume 136
4th International Conference on Sensors, Mechatronics and Automation (ICSMA 2016)

Research on Semi-supervised Classification with an Ensemble Strategy

Zhanhao Han, Shiqun Yin

Faculty of Computer and Information Science, Southwest University,Chongqing, China

731358416@qq.com,qqqq-qiong@163.com

Keywords: sentiment classification; semi-supervised learning; ensemble learning

Abstract. The classification approach based on semi-supervised learning, which is based on small sample sizes marked on the sample by means of a non-labeled improve the classification performance. In order to improve the ability of semi-supervised learning, this paper presents an approach based on the basis of the consistency of the label, the integration of two mainstream semi-supervised classification method is used: Collaborative training methods and label propagation method based on random feature subspace. First, the two semi-supervised learning methods to train the classifier to label unlabeled samples; secondly, select the marked sample rate without labels, so as to obtain a semi-supervised learning method is better than any of the classification results.

Introduction

Supervised learning requires a large number of labeled training samples, which makes the supervised-learning classification system need some manual labeling and time cost [1]. There is still a certain gap between the classification results and the practical requirements of unsupervised-learning approaches [2].

The specific implementation process [3], we selected the improved collaborative training algorithm [4] and label propagation algorithm [5] as the basic algorithm of semi supervised learning, and in the above two algorithms to generate classifiers for unlabeled samples are labeled after selecting two tagging consistent samples to the labeled sample concentration training classification model [6]. The label propagation algorithm(LP) assumes that the two nodes with the same characteristics tend to belong to the same category [7], using a small amount of labeled data to assist the unsupervised learning of large amounts of unlabeled data [8]. The greater the weight of the edge between the nodes [9], the easier the label information is transferred between nodes.

General framework for semi-supervised ensemble learning

Semi-supervised ensemble learning is a kind of learning mechanism which combines many kinds of semi-supervised learning methods. In the case of a given labeled samples and unlabeled samples, different semi-supervised learning approaches can be used to form a new semi-supervised learning method based on ensemble learning. Generally speaking, the difference is one of the important factors that affect the performance of ensemble learning. The greater the difference in the approach of integration, the more obvious the performance of ensemble learning will be. Therefore, we choose different semi-supervised learning approaches to integrate learning. The following is a general model of ensemble learning algorithm.

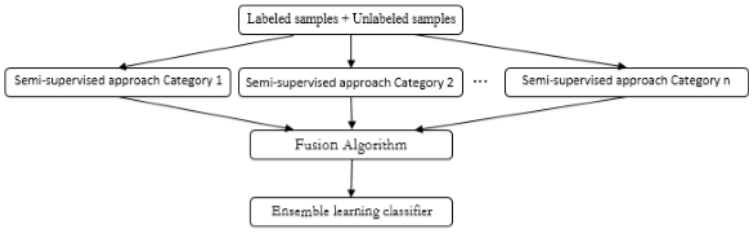

Fig1.general framework of semi-supervised ensemble learning

Semi-supervised classification approach based on Ensemble Learning

Co-training algorithm requires two independent views to train two classifiers, and then use the mutual aid approach to iteratively expand the data set and re-training. In order to obtain two independent views, an approach based on stochastic dynamic subspace is proposed for generating two different feature subspaces. In the concrete implementation, co-training algorithm of dynamic random subspace based on the generation, is each feature sub space as a representation of the view of text representation, different views of multiple feature subspaces corresponding to a plurality of text representation.

Used to update the labeled samples to classify the test samples.

Input:

Initial sample collection L has been marked, including n^+ positive class samples and n^- negative class samples;

Unlabeled sample collection U;

Output:

Updated labeled sample set L;

Program:

(1)B=Ø, B represents the final selected annotation consistent with the sample collection;

(2)Using different semi-supervised learning algorithm$F_i(i = 1,2)$ for each sample U in the X to mark the results for $L_i(X)(L_i(X) = c_1, \dots c_m)$;

(3)In order to take out each of the samples X in U:

If $L_1(X) = L_2(X) = c$

Unlabeled samples are labeled as class c and X is added to the B;

(4)Add the B to the labeled sample$(L = L \cup B)$ and remove it from the U.

We define two functions, in which X is the semi-supervised learning algorithm $F_i(X)$and the classification results of the unlabeled samples X:

$$E_1(\mathrm{X}) = \begin{cases} 1, L_1(X) = L_2(X) \\ 0, \ L_1(X) \neq L_2(X) \end{cases} \quad (1)$$

$$\mathrm{E}_2(\mathrm{X}) = \begin{cases} 1 \ \text{Consistent annotation sample X is correctly classified} \\ 0 \ \text{Consistency labeling sample X is incorrectly classified} \end{cases} \quad (2)$$

According to the above algorithm process, we can get the formula (3).

$$P(E_2(\mathrm{X}) = 1) = P(L_{ES}(\mathrm{X}) = real(\mathrm{X}) | E_1(\mathrm{X}) = 1) \quad (3)$$

Among them,$L_{ES}(X)$ for the integrated learning system for unlabeled samples (that is, the choice of sub semi-supervised learning approach consistency of tagging),$real(\mathrm{X})$ for the sample X real label.

Among them,$E_1(X) = 1$ and $L_{ES}(X) = c$ (c for a category)

Currently only when

$$L_1(X) = L_2(X) = c. \quad (4)$$

Each sub-supervised learning algorithm F_i on the classification accuracy of the unlabeled samples, but also can be used to correctly classify each unlabeled sample rate, that is,$P(L_i(\mathrm{X}) = real(\mathrm{X}))$, a simple representation for theP_i.

$$P\big(L_{ES}(\mathrm{X}) = real(\mathrm{X})\big) = P_1 \times P_2 \quad (5)$$

$$P(E_1(X) = 1) = P\big(L_1(X) = L_2(X) = real(X)\big) + P(L_1(X) = L_2(X) \neq real(X)) \quad (6)$$

Because the results of the sample classification are only two categories, Positive sample and negative sample , we can also be the type (5).

$$P(L_1(\mathrm{X}) = L_2(\mathrm{X}) \neq \mathrm{real}(\mathrm{X}) = (1 - P_1)(1 - P_2)) \quad (7)$$

Comprehensive (3), (5), (6), (7) the formula (8).

$$P(E_2(X) = 1) = \frac{P_1 \times P_2}{P_1 \times P_2 + (1 - P_1)(1 - P_2)} \quad (8)$$

Let $P_{best} = \mathrm{MAX}$, in order to prove the semi-supervised sentiment classification method based on ensemble learning reduced sub semi- supervised learning approach of the error marker is the sample rate, with marked higher accuracy, we need to prove (9).

$$\frac{P(E_2(X)=1)}{P_{best}} > 1 \quad (9)$$

We might as well assume that $P_{best} = P_1$, then (7) and can be written in formula (9).

$$\frac{P_2}{P_1 \times P_2 + (1-P_1)(1-P_2)} > 1 \quad (10)$$

Further simplified formula (9)

$$\frac{P_2(1-P_1)-(1-P_2)(1-P_1)}{P_1 \times P_2 + (1-P_1)(1-P_2)} > 0 \quad (11)$$

Experiment

Experimental setup. The experimental data includes :four areas of product reviews, including Books, DVD, Electronics and Kitchen in four different product reviews and film (Movie) domain corpus. Each domain contains 1000 positive and 1000 negative reviews. Two different experimental settings: (1)Randomly selected 5% samples as the initial marked samples, 85% samples as unlabeled samples; (2) Randomly selected 10% samples as the initial marked samples, 80% samples as unlabeled samples;

Table 1

	Prediction for positive class samples	Prediction for negative class samples
Positive class samples	Correct positive class samples(TP)	Class samples for classification errors(FN)
Negative class samples	Negative class samples for classification errors(FP)	Classification of negative class samples(TN)

Experimental results and Analysis. Fig. 2 shows the classification performance of various semi-supervised learning approaches when the initial annotation sample is 5%.Fig.3 shows the classification performance of various semi- supervised learning methods when the initial annotation sample is 10%. The classification accuracy is 2.1% and 3.8% respectively compared with the Co-training algorithm and LP algorithm.

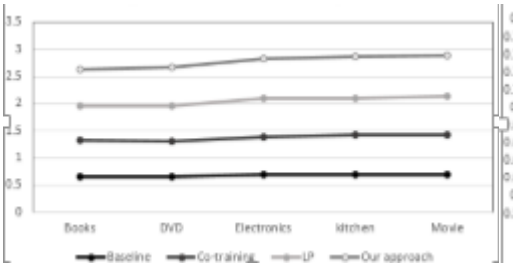

Fig2. initial annotation sample 5%

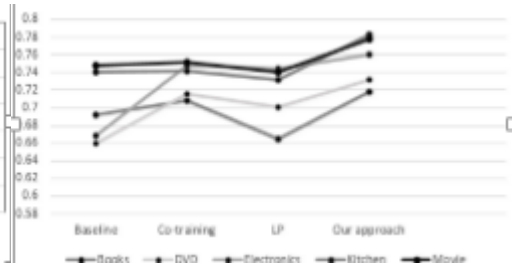

Fig3.initial annotation sample 10%

Table2 Initial annotation sample 5%

Field	Co-training	LP	Our approach
Book	0.29	0.43	0.19
DVD	0.29	0.42	0.18
Electronic	0.28	0.38	0.15
Kitchen	0.25	0.38	0.12
Movie	0.25	0.36	0.12

Table3 Initial annotation 10%

Field	Co-training	LP	Our approach
Book	0.25	0.41	0.12
DVD	0.26	0.39	0.11
Electronic	0.23	0.36	0.10
Kitchen	0.23	0.35	0.10
Movie	0.22	0.32	0.08

Conclusion

In this paper, we study the semi-supervised classification based on ensemble learning, and propose a semi-supervised ensemble learning method based on the uniform label fusion for positive and negative sample classification [11]. The approach can reduce the error rate of the semi-supervised learning algorithm to the unlabeled samples, and has a certain noise filtering function. Experimental results show that our approach can further improve the classification accuracy of semi-supervised sentiment classification [12], and the performance is better than the semi-supervised classification approach.

Acknowledgment

This work is supported by the Science &Technology project (2013001287， 41008114, and 41011215). Corresponding author: Shiqun Yin, qqqq-qiong@163.com.

References

[1] Li S,C Huang, G Zhou, et al. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Calssification [C]//Proceedings of ACL-10,(2015).

[2] Zhu X. and Z. Ghahramani..Learning from Labeled and Unlabeled Data with Label Propagation. CMU-CALD-02-107(2016).

[3] Blum A, T Mitchell. Combining Labeled and Unlabeled Data with Co-training [C]//Proceeding of COLT-98, (2015).

[4] Pang B.L Lee, S Vaithyanathan. Thumbs up? Sentiment [C]//Proceedings of EMNLP-02,(2012).

[5] Zhao SB; Grishman R Extracting relations with integrated information using kernel methods (2013).

[6] Chen JX; Ji DH; Chew LT; Niu ZY Automatic relations among named entities from Large corpora (2012).

[7] He TT; Xu C; Feng YY; Huang RH Chinese automatic entity relation extraction-2007(04)

[8] Liu KB; Li F; Han Y Implementation of a kernel-based Chinese relation extraction system-2014(08)

[9] Zhang SX; Wen J; Qin Y; Yuan CX, Zhong YX Study about automatic entity relation extraction-Journal of Harbin Engineering University 2014(B07)

[10] Banerjee S, Ramanathan K, Gupta A. Clustering Short Texts Using Wikipedia[C](2015).

[11] Day N E. Estimating the Components of a Mixture of Normal Distributions [J](2014).

[12] Dempster A P, Laird N M,Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J](2015).