Network Video Online Semi-supervised Classification Algorithm Based on Multiple View Co-training

Nanbin Zhang and Xiaoping Liu
Chongqing Three Gorges Vocational College, Wanzhou 404155, China
E-mail: znb1010@163.com, lxp928@163.com

Abstract—As information integration based on multiple modal has to problems like complexity calculation process and low classification accuracy towards network video classification algorithm, came up with a network video online semi-supervised classification algorithm based on multiple view co-training. According to extract the features in text view and visual view, to the feature vector in each view, uses graph as basic classifier and modeling, uses linear neighborhood belief propagation to make category labels propagation in each view, and gets category prediction outcomes in this view; in different views, uses co-training method to online extract unlabeled samples to expand the training set and to incrementally update basic classifier. To the integration of different model prediction outcomes, proposed an integration method aimed at category related. Finally made detailed experimental compare with support vector machine, the performance of learner increased greatly, more suitable for large-scaled network video online semi-supervised learning.

Keywords—incremental online learning; text view; visual view; multiple model integration

I. INTRODUCTION

Network video is a kind of significant data in the Internet, has the characteristics as large scale of data, heterogeneous information coexistence and many unmarked data [1]. Integrated uses heterogeneous information like visual, text and so on, is very important to make online semi-supervised classification for network video [2-4].

In recent years, many scholars used information integration based on multiple modal and made researches on network video classification, Yang Lin jun, etc, used the information like bottom-layered features, semantic features, audio features, additional text features and so on to classify, the result showed, multiple modal classification result was better than single modal, and the effect of support vector machine was the best [5-8]. Cui Bin, etc, extracted the visual features in training set to assist the calculation of word similarity in text message, avoided the extraction of visual feature in classification, increased the velocity and performance of classification [9-11]. Zhang Xu, etc, made characterizing definition used semantic concept model, used the semantic information contained in the text to increase the classification performance, and used incremental support vector machine to classify, but the calculation process was relatively complex [12-15]. To make better use of the relationship among data, Wu Xiao, etc, integrated used network video’s header and label, relative video information, video uploader’s personal preference and so on to improve the classification performance. Chen Zhi neng, etc, used the result feedback by search engine to assist network video classification [16-17]. The works all above place emphasis on supervised classification but cannot efficiently make online model update. The structure of this text is: the second part in introduced the system of network video online semi-supervised classification, the third part introduced the online semi-supervised classification model and detailed description of the online semi-supervised classification algorithm, finally, made a experimental simulation of the online video semi-supervised classification algorithm based on multiple view co-training.

This paper mainly made explosive and innovative work at the following aspects:

Aimed at the work of multiple modal information integration towards network video classification, placed emphasis on supervised classification but not online semi-supervised classification, cannot efficiently make online model updating, came up with the network video online semi-supervised classification algorithm based on multiple view co-training. Firstly extracted the features in text and visual, to the feature vector in each view, took views as the basis classifier and modeled, and used linear neighborhood propagation to make propagation of category labels of each view, then got the category predicting result in this view; in different views, online extracted unlabeled samples used co-training method to expanding the training set and incrementally updated the basis classifier. To the integration of different modal’s prediction results, came up with a integration method aimed at category related. This method used local learning method to model and learns of the data, could get a better accuracy than global learning.

In order to further prove the accuracy and efficiency of network video online semi-supervised classification algorithm based on multiple view co-training, made
contrast experiment compared with the classification algorithm of support vector machine, made performance qualification to the category specific integration and the weight unified integration, the experiment result showed: on the condition of a small amount of labeled samples, the method in this text in text view was obviously superior to SVM algorithm, in visual view the two was approximately the same. After the two views integrated, the method in this text was superior to SVM algorithm about 8.3%; the overall classification accuracy of this algorithm reached 99%. The online network flow classification and the accuracy of classification prediction greatly increased, which fitted well the large-scaled online learning.

II. PROPOSED SCHEME

As shown in Figure 1, this frame contains multiple modal feature extraction and online semi-supervised classification. The multiple modal feature extraction mainly contains the extraction to text features and visual features; in online semi-supervised classification, each modal is regarded as a view and uses graph as basic classifier, and then online update the learner according to co-training method label and extracts unlabeled samples. The specific steps is as following: (a) in initial training set, uses linear neighborhood propagation to learn about the basic classification; (b) in the process of online learning, according to co-training method, uses text (visual) view to predict the unlabeled samples, extracts the samples with high confidence coefficient and with representativeness applies to the visual (text) view; (c) uses learned classifier to classify the text sample. In actual application, the two processes of online learning and classification can operate at the same time.

A. Online Semi-supervised Classification Model

In each view, takes sample as apex, the similarity among samples as rims and makes the graph \( G=(V,W) \), there \( V \) contains initial i labeled samples: \( L=\{x_1,y_1\}, \ldots, (x_i,y_i) \}, \) and \( u \) unlabeled samples, \( U=\{x_{i+1}, \ldots, x_{i+u}\} \}, \) and the dimension is e of input sample \( x_i \subset R^e \); category label vector \( y_i=[y_{i,c}]_{c=1,c}^C \), if \( x_i \) belongs to c category, \( y_{i,c}=1 \) and \( y_{i,c}=0(c \neq c) \)

Supposes each point \( x_i \) in the graph can be linear reconstruction by K neighborhood point \( x_{ij} \in N(x_i) \). The reconstruction weigh coefficient can be got by minimum error \( \delta_i \):

\[
\delta_i = \|x_i - \sum_{j=1}^K n(x_i)w_{ij}x_{ij}\| \\
\text{s.t.} \sum_{j=1}^K n(x_i)w_{ij} = 1, w_{ij} \geq 0
\]

In the formula: \( w_{ij} \) is reconstruction weight coefficient. Weight vector is \( w_i=[w_{i1},w_{i2},...,w_{ik}]^T \), weight matrix is \( w=[w_{ij}]_{i,j} \). The predicting value e of linear neighborhood propagation calculation by iteration input sample:

\[
e_i = \sum_{j=1}^K n(x_i)w_{ij}y_{ij}^{t-1}
\]

In the formula, \( e_i^{t-1} \) refers to the prediction value of \( x_j \) when t-1 times of iteration. Decomposes W as:

\[
R=[R_1,R_2,R_3,R_4]
\]

In the formula, \( R_1, R_2, R_3 \) and \( R_4 \) are four matrix. The prediction outcomes 0.0 in U convergent to \( R=(1-R_{tt})^{-1}R_p \)

In the formula, \( p=[p_c]_{c=1} \) is label matrix.

In online learning, after the training set \( x_0 = L, u \) adding new samples every time, formula (2) should be recalculated in all samples, the calculation complexity is high and unsuitable for online learning. So the weight vector \( w_i \) is revised as following:

\[
\hat{w}_i = \phi w_i, \phi = \text{diag}(\theta_1, \theta_2, ..., \theta_d)
\]

In the formula: \( \hat{w}_i \) is respectively the weight vector before and after revise, \( \theta_\gamma = \exp(d_{ij} - d_{max})^2 / \sigma^2 \), \( d_{ij} \) is the distance to its neighborhood point \( x_j, d_{min}, d_{max}, d_{ij} \) and \( \sigma^2 \) is the variance of sample distance \( \{d_{ij}\} \), this text values \( \sigma = 10 \).

From formula (2) and (4), category label propagation and decay feature and brings the advantages as following: label prediction can expresses confidence coefficient, limits the noise brings by the error-labeled samples at a local area, and in favor of realizing the incremental update of model. Supposes the dataset only contains one label sample \( (x, y) \), the unlabeled sample according to label propagation \( \hat{f}_i = [f_{i1}, f_{i2}, ..., f_{iu}] \) can be known by formula (2) and (4):

\[
f_i = \sum_{\hat{w}_{ij} \neq 0} \max f_{ij} \sum_{\theta_j} w_{ij} = \hat{\xi}_i f_i
\]

In the formula:

\[
\hat{\xi}_i = \sum_\theta w_{ij} \cdot 1, f_i^* = \max f_i
\]

Repeats formula (5) can get the first node is \( x_{u(1)} = x \)
in the path \( s_i \) from \( x \) to \( x_j \), the sequence number of \( m \) node is \( s(m) = \arg \max_{x_{ij} \in N(x(i)(m-1))} f_{x_{ij}} \), in this path, the following inequality is true:
\[
f_{x_{ij}} \leq y_c \prod_{k=1}^{m} \xi_{ik} \leq y_c (\max y_k)_{1} \leq y_c (\max y_k)_{1}
\] (6)

In this formula: \( s_i \) is the shortest path from \( x \) to \( x_j \), \( |x_j| \) refers to the distance of path \( s_i \), \( y_c \) is the \( c \) component of \( y \). Formula (6) expresses label propagation takes exponential function as its upper bound, limits label propagation in local area \( L_d(x) \) around \( x \), when training set increases new labeled sample \( (x_j,y_j) \), this text makes incremental update using following formula:
\[
f_{x_{ij}}[f_{x_{ij}},f_x] + f_{x_{ij}}
\] (7)

In the formula:
\[
f_{x_{ij}} = (1 - w_{in})^{-1} w_{in} y_{c}^{-v}
\]
And calculates only in \( L_d(x) \), \( f_x \) is the prediction to \( x \) based training set \( x \), \( x_{ij} \) at time \( t-1 \).
\[
f_{x} = \sum_{ix_{ij}\in L_d(x)} w_{in} f_{x_{ij}}
\]
The characteristic of algorithm is: when training set added into new labeled sample, the label prediction only needs to make incremental calculation in \( L_d(x) \). Figure 2 shows the calculated result in incremental update and all data is quite similar.

Online semi-supervised classification

Network video online semi-supervised classification algorithm processes based on co-training frame. The collected unlabeled sample set during online operating process is setted as Up. The online semi-supervised learning process respectively predicts and labels the sample in Up according to the basic classifier in text and visual view; according to prediction result, extracts “good” unlabeled sample and its predicting label and adds into the opposite training set; makes basic classifier incremental update in new training set

a) The category label prediction in Up, given \( x \), \( U \), the label prediction value of view \( \nu = (1,2) \) is:
\[
f_{\nu}^{(x)} = \sum_{i} w_{i} f_{\nu}^{(x)}
\] (8)

In the formula, \( x_{ij} \) is the neighborhood point of \( x \) in training set \( X_i \). According to formula (8), the calculation need not change the structure of whole graph thus decreases the complexity of calculation.

b) The extraction of unlabeled sample

How to extracts unlabeled sample to expand training set is the essential problem of semi-supervised learning. The “good” unlabeled sample should be: (1) high confidence coefficient of label prediction, the bringing of error prediction sample is equal to brings high noise; (2) reasonable distribution in feature space, if the added unlabeled sample cannot better represent the whole sample distribution, the classifier preference improvement would be slow and has bad expansiveness.

From formula (6), if \( f_{\nu}^{(x)} \) in formula (8) has bigger component than \( f_{\nu}^{(x)} \), it shows the sample is close to labeled sample. According to manifold assumption, the category of \( x \) is similar to sample \( x \), so the label prediction result in formula (8) contains the confidence coefficient of label prediction result. For example, \( f_{\nu}^{(x)} = 0.96 \) shows sample \( x \) has high proportion of being judged as \( c \) category in view \( v \). On the other hand, the sample with big \( f_{\nu}^{(x)} \) always close to labeled sample, so extracts certain biggest prediction samples from Up would not benefit for category information quickly expanding to the whole space. According to those analysis, extracts “good” sample in view 1 and offers to view 2 by following ways: orders from the biggest to the smallest of the prediction value \( f_{\nu}^{(x)} \) and \( f_{\nu}^{(2)} \) in Up, values \( f_{\nu}^{(x)} \) is the biggest and \( f_{\nu}^{(2)} \) not in the first 5% \( m_u \) sample. Figure 3 is the schematic diagram of sample extraction, there the black and white node is the samples in \( LxUUs \), grey node \( X_1^{(1)}, X_2^{(1)}, X_3^{(1)}, X_4^{(1)} \) is the node in . In view 1, the prediction value is pretty big but in view 2 the predicting value is big. From geometric angle, brings \( X_4^{(2)} \) in view 2 can actually get makes training set’s distribution more suitable for the sample’s real distribution than brings \( X_5^{(2)} \). The “good” sample extracts in view 2 has similar method of extracting for view1.
views of training data $X_{tr}$ calculates rim weight, makes prediction for sample category label according to linear neighborhood propagation, gets basic classifier $G_1$ and $G_2$;

c) Online gets p video data $U_p = \{x_1, ..., x_p\}$, respectively calculates the text and visual features’ vector sets, sets as $u^{(i)}_x$ and $u^{(i)}_v$,

d) Uses $G_1$ to predict $u_p^{(i)}$ and gets the result $f_p^{(i)}(1) = [f^{(1)}_1, f^{(2)}_1, ..., f^{(c)}_1]$ uses $G_2$ to predict $u_p^{(i)}$ and gets the label $f_p^{(i)}(2) = [f^{(1)}_2, f^{(2)}_2, ..., f^{(c)}_2]$ meanwhile, according to the category label in chapter 2.4 into tomorrow (65/号) 中午 12 点半在我办公室开+(6) of model updating in each view; in $L_{rk}$(*), makes incremental updating basic classifier according to formula (7).

Category related multiple modal integration

Different modals have different classification abilities, the modal with strong ability of classification has relatively high predicting confidence coefficient to category label. Meanwhile, the same modal has different classification ability to different category data. So, integration weight coefficient should be category related. This text confirms this coefficient based on $F_1$, $F_{1}=2pr/(p+r)$, there $p$ and $r$ are respectively refers to the accuracy and recall rate.

Given validation set $V=(V_1, V_2)$, $V_1$ and $V_2$ respectively corresponds to text view and visual view. Makes prediction of $V_1$ and $V_2$ by basic classifier $G_1$ and $G_2$, the result are $F_{1} = (F_1^1, F_1^2, ..., F_1^{c_1})$, $F_{2} = (F_2^1, F_2^2, ..., F_2^{c_2})$, $d_{ij}$, $f_i^1$s and $f_i^2$s are respectively the value of $F_1$ matches by $i$ category classifies by $G_1$ and $G_2$. Sets the weight coefficient vector of $G_1$ and $G_2$ are respectively $w_i = [w_i^1, w_i^2, ..., w_i^{c_1}]$ and $w_v = [w_v^1, w_v^2, ..., w_v^{c_2}]$.

This text confirms this weight coefficient by the following formula

$$w_i^1 = \frac{f_i^1}{f_i^1 + f_v^1}, \quad w_i^2 = \frac{f_i^2}{f_i^1 + f_v^2}$$

And makes the modal with high classification accuracy accounts big weight. For test sample $x_i$, if $G_i$ makes prediction for $U_p^{(i)}$ and gets the result as $\hat{Y}_i = (f^{(1)}_i, f^{(2)}_i, ..., f^{(c)}_i)$, the predicting result of $G_2$ is $\hat{Y}_i = (f^{(1)}_i, f^{(2)}_i, ..., f^{(c)}_i)$, the predicting result after integration is:

$$\hat{Y}_i = \left(\frac{w_i^1 f_i^{(1)} + w_i^2 f_i^{(2)}}{w_i^1 + w_i^2}, ..., \frac{w_i^1 f_i^{(1)} + w_i^2 f_i^{(2)}}{w_i^1 + w_i^2}\right)$$

The belonging category is:

$$\arg\max_{c \in C} \left(\frac{w_i^1 f_i^{(1)} + w_i^2 f_i^{(2)}}{w_i^1 + w_i^2}\right)$$

### III. EXPERIMENTAL RESULTS

#### A. Experiment Data Set

MCG-WEBV is the structured network video dataset based on Youtube, its visual features contains color histogram, color moments, edge histogram descriptor and so no. The text message is expressed by bag of words model. This text extracts 7 categories, 29437 network videos as experiment data. The detailed information shows in table 1.

<table>
<thead>
<tr>
<th>Number</th>
<th>Video category</th>
<th>Sample size</th>
<th>Scale(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MUSIC</td>
<td>6897</td>
<td>24.09</td>
</tr>
<tr>
<td>2</td>
<td>Entertainment</td>
<td>5187</td>
<td>18.45</td>
</tr>
<tr>
<td>3</td>
<td>News and Politics</td>
<td>4713</td>
<td>15.91</td>
</tr>
<tr>
<td>4</td>
<td>Sports</td>
<td>3980</td>
<td>13.26</td>
</tr>
<tr>
<td>5</td>
<td>Pets Animals</td>
<td>3210</td>
<td>11.02</td>
</tr>
<tr>
<td>6</td>
<td>Autos Vehicles</td>
<td>3209</td>
<td>10.89</td>
</tr>
<tr>
<td>7</td>
<td>Gaming</td>
<td>2348</td>
<td>7.97</td>
</tr>
</tbody>
</table>

#### B. Classification Results

This chapter makes experimental compare of SVM and the method in this text in the environment of less sample. The data divided into labeled sample set and unlabeled sample set, the unlabeled sample set has 2964 data, the scale of labeled sample set are 70, 140, 210, 280 and 350. The two method both training in labeled sample set, predicting in unlabeled sample set. SVM realizes based on Libsvm toolkit, to text data, uses linear kernel, to visual data, uses radial basis function kernel function, extracts parameter to make classification performance the best. According to experiment, local neighborhood $L_k(*)$ is 3.5K. When forms the graph based on formula (1), text view takes included angle cosine as distance measure, visual view takes euclidean distance. The experiment result shows as figure 4. From the result, in the environment of less labeled sample, the text method performance in text view is greatly better than SVM algorithm, in visual view, the two performance are similar. After the two views integration, the method in this text is better than SVM algorithm about 8.3%.
In this paper, methods SVM Classification accuracy/%

Figure 4. The classification performance's compare result of the method in this text and SVM

C. Online Learning Accuracy

This text mainly analyzes on the classifier online updating effect on classification performance and unlabeled sample extraction method’s effect on classification performance.

This experiment compares the classifier under the online updating and off-line no updating to online data’s prediction performance. Divides experiment data by months into training data and online data. Training data package contains the data in the first and second months, online data package contains all the data in the third to the eighth months, there the online data of each month randomly divides a part using as test set, to test the prediction performance of the classifier to the month (i.e. online data) in learning. Online learning 80 times, the experiment result showed as figure 5.

From the experiment result, classifier online updating has little differences with off-line on updating in first several months keeps the same in certain degree, with time goes, the performance differences of the two is obvious in later several months. For the performance change trade, the classifier online updating gradually expresses better classification performance later.

b. Unlabeled sample extraction method’s effect on classification performance

In multiple view co-training, how to predict unlabeled sample label in multiple view and to extract unlabeled sample to expand training set has significant effect on semi-supervised learning performance. This experiment analyzes this following three method’s performance effect on online learning: when online learning, randomly increase labeled sample into sample set; when online learning, extracts unlabeled sample with high predicting confidence coefficient and its predicting label to expand training set; when online learning, randomly extracts the training set of unlabeled sample. The second method is the online learning method in this text, the first method is equal to supervised learning method, the third method is equal to the classifier totally not predict, randomly extracts unlabeled sample to expand training set.

Figure 6. Performance change of the three sample extraction method under multiple modal integration

In the experiment, online learning 80 times, the result is shown as in figure 6. At the beginning of online learning, because the classifier and training data under each method has little differences, so the performance differences is little; with online updating processing, the differences of classifier’s performance gradually becomes bigger. From figure 5, the performance of the second method among the first and the third, in m5 and m6, method 2 increases a lot compares with method 3, this shows the method in this text can online increase the performance of classifier. The increase of performance has close relationship with basic classifier: if the basic classifier has high accuracy and good generalization
performance, then online semi-supervised learning has good performance improvement.

D. Category Related Multiple Model Experiment

The experiment in this chapter compares the result in similar weight coefficient and category related weight coefficient. The experiment uses 6157 training set, which includes 2894 labeled sample and 3263 unlabeled sample, uses for calculating specific categories’ integrate weight coefficient testing set contains 2895 labeled sample, 24842 sample in test set. Calculates in test set and specific categories combine weight coefficient and integrate result shows in table 3. Category unified integrate weight confirms by search on grid (10%, 20%, 60%, 90%), makes the classification performance is the highest after integration, finally makes sure visual and text modal makes prediction, extraction, inter-label and expansion training set to online unlabeled sample by co-training method, makes the classifier can learn online and keeps unification with online data distribution. The multiple view co-training has following advantages: (a) As local learning always has better effect than global learning, so the method in this text can get better result; (b) Uses local learning into co-training can simply and efficiently makes incremental learning, pretty suitable for large-scale date’s online learning. The experiment result shows, in the environment of less labeled sample, the method in this text is better than wildly-used SVM algorithm, compares it with the environment of classifier off-line no updating, classifier online learning gradually shows the increase on performance. Besides, specific category multiple information integration realizes more efficiently multiple modal information integration.

| TABLE 2. COMPARE OF SPECIFIC CATEGORY INTEGRATE WITH UNIFY WEIGHT INTEGRATION PERFORMANCE |
|-----------------------------------|------------------------|------------------|------------------|----------------------|
| Video category                   | Categories of uniform weight | Category specific weight | |
|                                  | F_1            | F_2            | w_1             | w_2             | F_3             |
| Music                            | 60             | 60             | 35.98           | 51.32           | 49.80           | 37.03           |
| Entertainment                    | 60             | 60             | 54.09           | 54.23           | 47.34           | 54.04           |
| News                              | 60             | 60             | 41.23           | 59.08           | 53.78           | 53.45           |
| Politics                         | 60             | 60             | 60              | 55.43           | 64.23           | 64.09           |
| Sports                           | 60             | 60             | 43.05           | 56.21           | 44.23           | 45.23           |
| Autos                            | 60             | 60             | 65.02           | 59.45           | 42.64           | 65.04           |
| Vehicles                         | 60             | 60             | 46.69           | 53.04           | 48.07           | 47.79           |
| Pets                             | 60             | 60             | 49.87           | 55.02           | 46.34           | 50.34           |

W_1 and W_2 in table 2 are respectively weight of text and visual view. From the experiment result, the result of the two integrated method has large improvement than single view’s classification result (seen in the last line in table 3), the performance increase range all higher than 10%. Secondly, specific category weight integrate method has certain improvement than unify weight, from table 2, the former higher 0.8% than the later. Besides, the result shows text modal has better classification ability than visual modal, accounts more weight in integration. In realistic application, specific category weight’s integrate method needs the predicting outcomes of single modal to extracts possible several sets of weight, finally extracts the category with the biggest possibility after integrating as final result.

IV. CONCLUSION

This paper proposed multiple view co-training method to solve the problems of multiple modal network video online semi-supervised classification. This model takes view as basic classifier and uses linear neighborhood propagation with decay to predict sample. In this text, the classifier corresponding to visual and text modal makes prediction, extraction, inter-label and expansion training set to online unlabeled sample by co-training method, makes the classifier can learn online and keeps unification with online data distribution. The multiple view co-training has following advantages: (a) As local learning always has better effect than global learning, so the method in this text can get better result; (b) Uses local learning into co-training can simply and efficiently makes incremental learning, pretty suitable for large-scale date’s online learning. The experiment result shows, in the environment of less labeled sample, the method in this text is better than wildly-used SVM algorithm, compares it with the environment of classifier off-line no updating, classifier online learning gradually shows the increase on performance. Besides, specific category multiple information integration realizes more efficiently multiple modal information integration.

REFERENCES

[12] Kuo-Feng Huang, Shih-Jung Wu. Real-time-service-based Distributed Scheduling Scheme for IEEE 802.16j


Zhang Nanbin, born in NanChong Sichuan in 1969 male, Han, master of engineering, associate professor, main research direction: computer application technology.

Liu Xiaoping, born in Chongqing Wanzhou in 1971, female, Han, bachelor's degree, senior engineer, main research direction: computer control technology.