Vol.2015-CVIM-198 No.16
2015/9/15

MRF-Based Multi-Label Classification using Label Relations

Ryosuke Furuta[1,a)] Yusuke Fukushima[1,b)] Toshihiko Yamasaki[1,c)] Kiyoharu Aizawa[1,d)]

Abstract: Multi-label classification and multi-classifier fusion have been independently explored as different problems. We propose a re-labeling method that can simultaneously treat these two problems in an unified framework. Our method considers (a) correlations between different labels, and (b) correlations between different feature types. In particular, the proposed method models both label and feature correlations in a single Markov random field (MRF), and jointly optimizes the label assignment problem. We apply our method to impression prediction of oral presentations. We train and evaluate the proposed method using a collection of 1,646 TED talk videos for 14 different impression types. Experimental results on this dataset show that the proposed method obtains a statistically significant macro-average accuracy of 93.3%, outperforming several competitive baseline methods.

1. Introduction

Multi-label classification and multi-classifier fusion are important tasks in the field of machine learning. The former is a task where multiple binary labels ("yes" or "no") are assigned to input data for each pre-defined class. The latter is a task where scores by different classifiers are combined to obtain a final classification result. Although these two problems have long been explored, they have been independently treated as different problems so far. Based on the assumption that more accurate solutions can be obtained by treating these two problems simultaneously, we propose, *re-labeling*, a method that can treat multi-label classification and multi-classifier fusion in an unified joint optimization framework. Our *re-labeling* method considers labels predicted by a set of classifiers for a particular test instance, and selects the best subset of labels such that the correlations among the labels in the training data are optimally satisfied. Specifically, we use quadratic pseudo boolean optimization (QPBO) [17] for this purpose. Simultaneously, our method can also treat *late feature fusion* that combines predictions from classifiers trained using different feature types to consider the correlations among features. In other words, our proposed method incorporates multiple features to predict multiple labels. First, we train several independent multi-label classifiers using different types of features. Second, we propose a Markov random field (MRF)-based label assignment method that considers (a) the relationships among different label types (i.e. *label correlation*), and (b) the relationships among different feature types (i.e. *feature correlation*) within a single joint optimization setting.

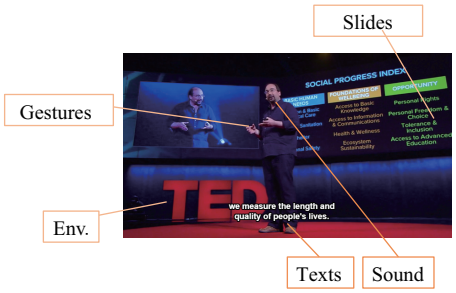


Fig. 1: Oral presentations are rich in multimedia data.

Although our method is general and can be applied to a variety of multi-label classification problems, in this paper we apply our method to predicting user impressions on a video presentation. In particular, the proposed method predicts *multiple* impression categories for a single presentation. In oral presentations, the impression labels assigned to a particular presentation are often highly correlated. Therefore, our method can successfully predict the impression labels by considering the *label correlations*. With regard to *feature correlation*, oral presentations are rich in multimedia data because they encompass a multitude of information sources such as visual aids/slides, careful control of voice (act of speech), selection of words, background effects, and physical gestures as shown in Fig. 1. A speaker will use gestures while referring to some text on a slide, while at the same time reading the text out loud. Because features that represent an oral presentation are not independent, an impression prediction method must consider the relationships among different types of features extracted from an oral presentation. Therefore, the impression prediction of oral presentations is also an ideal problem for multi-classifier fusion learned by different types of features such as linguistic and acoustic features.

Our algorithm has been successfully applied to 1,646 oral pre-

[1] The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[a)] furuta@hal.t.u-tokyo.ac.jp
[b)] fukushima@hal.t.u-tokyo.ac.jp
[c)] yamasaki@hal.t.u-tokyo.ac.jp
[d)] aizawa@hal.t.u-tokyo.ac.jp

sentations in TED Talks [2]. In particular, the proposed impression prediction method achieves a macro-averaged accuracy of 93.3% over 14 impression types. That of a baseline method that uses the same set of features as used by the proposed method but assumes all label types to be independent to train and test a multi-label classifier is 89.2%, that of the proposed late feature fusion is 90.5%, and that with the proposed relabeling is 91.7%. The improvements over those baselines are statistically significantly better ($\rho < 0.01$ between the proposed and the early feature fusion, and between the proposed and the late feature fusion, and $\rho < 0.05$ between the proposed and the relabeling, according to the Student's t-test.).

Our contributions can be summarized as follows.

(1) Our multi-label classification method considers the relationships between the labels by the MRF formulation, and produces a globally optimal label assignment for a given test instance. In particular, the proposed method does not depend on a particular set of labels, and can be applied in general to any multi-labeling task.

(2) We propose, late feature fusion, a method for considering the correlations among different feature types during label assignment. Both multi-label classification and late feature fusion is solved jointly using the same MRF. Although there are several multi-label classification methods and late feature fusion methods, to the best of our knowledge, they have not been used in a joint optimization task.

2. Related Work

2.1 Multi-Label Classification

There are two types of approaches in multi-label classification: classifier ensembles with problem transformations, and extensions to existing algorithms that can predict single labels.

In the first approach, each class is treated independently to train one-vs-rest binary classifiers. Next, the set of labels predicted by the individual classifiers for a test instance is considered as the set of labels for that test instance. An advantage of this approach is that existing binary classification algorithms can be readily used for multi-label classification. For example, off-the-shelf machine learning libraries such as LibSVM [7] and scikit-learn[*1] implement multi-label classifiers following this approach.

The second approach directly models the multi-label classification problem. For example, in [24], relationships between two labels in multi-labeled text categorization problem are elegantly formulated using the problematic generative models called parametric mixture model. However, this model can be used only when features are histograms of frequency such as Bag-of-Words. In [8], only exclusive relation is considered, and in [4], [5], only overlap and subsumption relations are considered. In [16], the propagation of the confidence scores of the assigned labels from training examples to test examples are formulated as a linear programming problem. In [27], k nearest neighbors to the test data are retrieved and the labels were determined by the maximum a posteriori (MAP) principle, and in [26], an error function capturing the characteristics of multi-label learning was proposed and it was incorporated into the back propagation scheme in the neural networks. In [10], the multi-labeling problem was solved by the loss minimization framework. In [15], [21] statistical topic models were introduced. A detailed comparison of 12 multi-label learning methods can be found in [19].

Our approach is inspired by [11], in which a formalism that captures three types of semantic relations (mutual exclusion, overlap, and subsumption) was proposed between any two labels applied to the same object. However, the semantic relations were manually defined, and the rules were hard constraints that handles only binary (i.e. *yes* or *no*) relations. In contrast, our proposed method conducts a soft assignment of multi-labels by taking into consideration the relationships among labels.

2.2 Late Feature Fusion

Feature fusion has been discussed in multimedia and computer vision communities because multimedia data are inherently multimodal, and different types of features can be obtained. There are mainly two types of feature fusion methods: *early fusion* and *late fusion*. In early fusion, multiple features are simply concatenated or merged by using dimensionality reduction or feature selection techniques before training the classifier. In late fusion, on the other hand, separate classifiers are trained using each feature type, and their outputs are somehow aggregated during test time. Late fusion has the advantage over early fusion that you do not have to determine which feature types must be concatenated before the training. Consequently, we focus only on late fusion.

In multiple kernel learning (MKL) [3], [13], the weighted sum of the individual decision values is calculated. The weights are trained in advance and held fixed during the test phase. In [18], sample-specific late fusion was proposed by formulating the dynamic weight allocation problem an $L\infty$ norm constrained optimization. On the other hand, likelihood ratio-based fusion [20] and a probabilistic framework [23] have been discussed where the labels from classifiers were considered instead of using the decision values.

3. Proposed Method

3.1 Re-Labaling using Label Relationships

3.1.1 Formulation

For ease of understanding, we describe the proposed method using the impression prediction of oral presentations as an example of multi-label classification problems.

Most methods proposed for multi-label classification assign each label independently using a binary classifier such as support-vector machine (SVM) trained separately to predict a single label. However, in oral presentations, the labels assigned to a particular presentation are often highly correlated. We refer to this phenomenon as *label correlation*. Therefore, it is appropriate to assign a set of labels that are coherent and consistent to an oral presentation. There is no guarantee that a set of independently trained classifiers would assign labels that are coherent to an oral presentation. To overcome this challenge, we propose a re-labeling method by leveraging the relationships between labels. Our proposed method simultaneously takes into account both the predictions by individual classifiers as well as the corre-

[*1] http://scikit-learn.org/stable/

lations among labels in training data.

Let $l_i \in \{0, 1\}$ be the binary label of the i-th impression, and $\boldsymbol{l} \in \{0, 1\}^n$ be the labeling of all impressions ($i = 1, \ldots, n$). Here, $l_i = 1$ means that the i-th impression is positive, and $l_i = 0$ corresponds to negative meaning (*e.g.*, if the 6-th ($i = 6$) impression is *informative*, $l_6 = 0$ means that the presentation is "not labeled as" *informative*). Because it is impossible to vote for "not" of the impressions, the impressions with less voting ratios were regarded as "not labeled." We model the labeling problem as a Markov random field (MRF):

$$E(\boldsymbol{l}) = \sum_i \phi_i(l_i) + \beta \sum_{i<j} \psi_{i,j}(l_i, l_j), \tag{1}$$

where ϕ_i is the unary term that represents how the i-th impression matches with the input presentation based on the decision of the i-th impression classifier, and $\psi_{i,j}$ is the pairwise term that represents the relationship between the i-th and the j-th impressions. β balances the unary and pairwise terms. By minimizing the Eq. (1), the optimal labeling that takes account of both the decisions of classifiers and the relationships between labels can be obtained.

We use the decision value of the i-th-impression classifier for the unary term ϕ_i, and use the sigmoid function

$$\varsigma_\alpha(x) = \frac{1}{1 + e^{-\alpha x}} \tag{2}$$

to convert the range of the decision values $(-\infty, \infty) \mapsto (0, 1)$. Therefore, the unary term ϕ_i is defined as:

$$\phi_i(l_i) = \begin{cases} \phi_i(1) = \varsigma_\alpha(-d_i) \\ \phi_i(0) = 1 - \varsigma_\alpha(-d_i), \end{cases} \tag{3}$$

where d_i represents the decision value of the i-th-impression classifier. When the decision value is positive ($d_i > 0$), the cost for $l_i = 1$ is lower than the cost for $l_i = 0$ (*i.e.*, $\phi_i(1) < \phi_i(0)$) and vice versa. The existing methods for multi-label classification can be regarded as a subset of our proposed model because each label l_i of the optimal labeling is independently decided by the sign of the decision value d_i where the energy function Eq. (1) has only unary terms (i.e. $\beta = 0$).

We define the pairwise term as follows:

$$\psi_{i,j}(l_i, l_j) = \begin{cases} \psi_{i,j}(0,0) = 1 - \frac{N_{ij}^{00}}{N_{ij}} \\ \psi_{i,j}(0,1) = 1 - \frac{N_{ij}^{01}}{N_{ij}} \\ \psi_{i,j}(1,0) = 1 - \frac{N_{ij}^{10}}{N_{ij}} \\ \psi_{i,j}(1,1) = 1 - \frac{N_{ij}^{11}}{N_{ij}}, \end{cases} \tag{4}$$

where N_{ij} is the number of training data that both the i-th and the j-th impressions are labeled. N_{ij}^{01} is the number of training data in which the i-th impression is labeled as 0 and the j-th impression is labeled as 1. The pairwise term Eq. (4) can be pre-calculated by counting the co-occurrences of each pair of labels in training datasets. Therefore, the more the number of co-occurrences of $l_i = 0$ and $l_j = 1$ in training datasets, the lower the cost $\psi_{i,j}(0, 1)$. In other words, we impose low costs on the pairs of labels that frequently co-occur in training datasets, and impose high costs on the pairs of labels that rarely co-occurs.

Next, we describe the derivation of Eq. (4). The objective here is to define the ideal pairwise term $\psi_{i,j}$ such that the optimal solution is as close to the true labels as possible for the test dataset. However, we of course cannot know the true labeling of the test dataset in advance. Therefore, the ideal pairwise term $\psi_{i,j}$ is predicted using the labels assigned to the training instances.

For the t-th training data ($t = 1, \cdots, N$), we first define the pairwise term such that the optimal solution is equal to the true labeling $\boldsymbol{l}^t = (l_1^t, \cdots, l_n^t)$:

$$\psi_{i,j}^t(l_i, l_j) \coloneqq \begin{cases} 0 & if\ (l_i, l_j) = (l_i^t, l_j^t) \\ \beta & else. \end{cases} \tag{5}$$

In Eq. (5), we impose a constant penalty β when the pair of labels (l_i, l_j) is not equal to the true labels (l_i^t, l_j^t). For example, when the true labels of the i-th and j-th impressions are 0 and 1 respectively (*i.e.*, $l_i^t = 0, l_j^t = 1$), the pairwise term $\psi_{i,j}^t$ is defined as follows:

$$\psi_{i,j}^t(l_i, l_j) \coloneqq \begin{cases} \psi_{i,j}^t(0,0) = \beta \\ \psi_{i,j}^t(0,1) = 0 \\ \psi_{i,j}^t(1,0) = \beta \\ \psi_{i,j}^t(1,1) = \beta. \end{cases} \tag{6}$$

We use the average of the Eq. (5) for all labeled training data in place of the ideal pairwise term:

$$\psi_{i,j}(l_i, l_j) \coloneqq \frac{1}{N_{ij}} \sum_t \psi_{i,j}^t(l_i, l_j). \tag{7}$$

Eq. (4) is obtained by deforming Eq. (7):

$$\psi_{i,j}(l_i, l_j) = \begin{cases} \psi_{i,j}(0,0) \\ \psi_{i,j}(0,1) \\ \psi_{i,j}(1,0) \\ \psi_{i,j}(1,1) \end{cases} \coloneqq \begin{cases} \frac{1}{N_{ij}} \sum_t \psi_{i,j}^t(0,0) \\ \frac{1}{N_{ij}} \sum_t \psi_{i,j}^t(0,1) \\ \frac{1}{N_{ij}} \sum_t \psi_{i,j}^t(1,0) \\ \frac{1}{N_{ij}} \sum_t \psi_{i,j}^t(1,1) \end{cases} \tag{8}$$

$$= \begin{cases} \frac{1}{N_{ij}} \beta (N_{ij}^{01} + N_{ij}^{10} + N_{ij}^{11}) \\ \frac{1}{N_{ij}} \beta (N_{ij}^{00} + N_{ij}^{10} + N_{ij}^{11}) \\ \frac{1}{N_{ij}} \beta (N_{ij}^{00} + N_{ij}^{01} + N_{ij}^{11}) \\ \frac{1}{N_{ij}} \beta (N_{ij}^{00} + N_{ij}^{01} + N_{ij}^{10}) \end{cases} = \begin{cases} \beta (1 - \frac{N_{ij}^{00}}{N_{ij}}) \\ \beta (1 - \frac{N_{ij}^{01}}{N_{ij}}) \\ \beta (1 - \frac{N_{ij}^{10}}{N_{ij}}) \\ \beta (1 - \frac{N_{ij}^{11}}{N_{ij}}) \end{cases} \tag{9}$$

$$(\because N_{ij}^{00} + N_{ij}^{01} + N_{ij}^{10} + N_{ij}^{11} = N_{ij})$$

β in Eq. (9) corresponds to β in Eq. (1).

Unlike [11] that can deal with only hard constraints such as "absolutely co-occurring" or "absolutely exclusive", our formulation can deal with soft constraints such as "tend to co-occur" or "tend to exclude". In addition, our method does not need prior knowledge about the relationships between labels, and can automatically predict them from training datasets.

In Fig. 2, we illustrate our MRF formulation as a graph structure. The energy function Eq. (1) corresponds to the undirected graph shown in Fig. 2. This graph has n nodes and each node corresponds to one impression. The label l_i is assigned to the i-th node ($i = 1, \cdots, n$). The i-th node has the unary cost $\phi_i(l_i)$ in Eq. (3). All pairs of the nodes are connected by edges. The edge

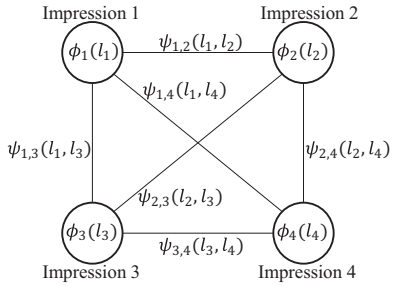


Fig. 2: The graph structure of Eq. (1) in the case where the number of the impression types is 4 ($n = 4$).

connecting the i-th and the j-th nodes corresponds to the pairwise cost $\psi_{i,j}(l_i, l_j)$ in Eq. (4) and represents the label relations between impressions.

3.1.2 Optimization

The global minimum of Eq. (1) can be obtained by graph cuts if the pairwise terms Eq. (4) are submodular:

$$\psi_{i,j}(0,1) + \psi_{i,j}(1,0) - \psi_{i,j}(0,0) - \psi_{i,j}(1,1) \geq 0. \tag{10}$$

However, the pairwise terms Eq. (4) are not always submodular because the values of pairwise terms depend on the training datasets. Therefore, we use QPBO method [17] to optimize the energy function Eq. (1). QPBO is a method optimized for binary labeling problems, which can exactly solve them even if the energy function is non-submodular. When the energy function is submodular, QPBO can obtain the same labeling as that of graph cuts (*i.e.*, global minimum). When the energy function is non-submodular, by allowing to assign "unknown" label $\emptyset$, QPBO can obtain a partial labeling of the global minimum: $\boldsymbol{l} \in \{0, 1, \emptyset\}^n$. For the cases where the output labeling of QPBO includes "unknown" labels $\emptyset$, a post-process to obtain a complete solution $\boldsymbol{l} \in \{0, 1\}^n$ has also been proposed (see [17]). Therefore, we can always obtain the global minimum of Eq. (1) by using QPBO.

3.2 Extension to Late Feature Fusion

Our MRF-based formulation described in Sec. 3.1 can be extended to late feature fusion. By the extension, our re-labeling method can simultaneously treat label relations between impressions and relations between the output labels from multiple classifiers learned by different features.

Let us consider the case where m classifiers are trained by different feature types for one impression (*i.e.*, there will be mn classifiers in total). Let d_i^p be the decision value of the classifier learned by the p-th feature ($p = \{1, \cdots, m\}$) for the i-th impression, and l_i^p be the label assigned to the i-th impression by the classifier learned by the p-th feature type. Let $\boldsymbol{l}' = (l_1^1, l_2^1, \cdots, l_{n-1}^m, l_n^m) \in \{0, 1\}^{mn}$ be the labeling by all mn classifiers. We re-formulate Eq. (1) by considering the multiple classifiers and the relations between them:

$$E(\boldsymbol{l}') = \sum_p \left(\sum_i \phi_i^p(l_i^p) + \beta \sum_{i<j} \psi_{i,j}(l_i^p, l_j^p) \right) + \sum_{p<q} \sum_i \varphi_{p,q}(l_i^p, l_i^q), \tag{11}$$

$$\phi_i^p(l_i^p) = \begin{cases} \phi_i^p(1) = \varsigma_\alpha(-d_i^p) \\ \phi_i^p(0) = 1 - \varsigma_\alpha(-d_i^p), \end{cases} \tag{12}$$

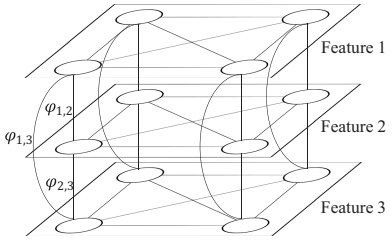


Fig. 3: The graph structure of Eq. (11) in the case when we use 3 classifiers for each impression ($m = 3$).

where pairwise term $\psi_{i,j}(l_i^p, l_j^q)$ is same as Eq. (4). $\varphi_{p,q}(l_i^p, l_i^q)$ is a new pairwise term between the output labels from multiple classifiers learned by the p-th and the q-th features. We define it as follows:

$$\varphi_{p,q}(l_i^p, l_i^q) = \begin{cases} \varphi_{p,q}(0,0) = 0 \\ \varphi_{p,q}(0,1) = \gamma \\ \varphi_{p,q}(1,0) = \gamma \\ \varphi_{p,q}(1,1) = 0. \end{cases} \tag{13}$$

In Eq. (13), γ is a constant penalty imposed when the i-th impression label from the classifier learned by the p-th feature and that by the q-th feature are different.

The energy function defined by Eq. (11) corresponds to the graph shown in Fig. 3. This graph has multiple m layers. In each layer there is a subgraph whose structure is same as in Fig. 2. The label l_i^p is assigned to the i-th node in the p-th layer. The p-th layer ($p = \{1, \cdots, m\}$) corresponds to the output labels from the classifiers learned by the p-th feature. The i-th node in the p-th layer has the unary cost $\phi_i^p(l_i^p)$ in Eq. (12). The intra-layer edges correspond to the pairwise costs $\psi_{i,j}(l_i^p, l_j^q)$ in Eq. (4) and represents the label relations between impressions. The interlayer edges correspond to the pairwise costs $\varphi_{p,q}(l_i^p, l_i^q)$ in Eq. (13) and represents the relations between the output labels from multiple classifiers learned by different features.

By minimizing Eq. (11), we obtain the output label set $\boldsymbol{l}' \in \{0, 1\}^{mn}$ that takes into account both the correlations between impressions and relations between the output labels from multiple classifiers learned by different features. However, the output label set $\boldsymbol{l}' \in \{0, 1\}^{mn}$ is redundantly long because our objective is to obtain n-dimensional labels $\boldsymbol{l} \in \{0, 1\}^n$. We solve the problem by setting the penalty γ to infinity ($\gamma = \infty$) in Eq. (13). When the penalty γ is extremely large, the output label set $\boldsymbol{l}'$ is forced to be $l_i^p = l_i^q$ ($\forall p, q$) by Eq. (13): Therefore, the output label set $\boldsymbol{l}' = (l_1^1, l_2^1, \cdots, l_{n-1}^m, l_n^m) \in \{0, 1\}^{mn}$ can degenerate into the n-dimensional labeling $\boldsymbol{l} = (l_1, \cdots, l_n) \in \{0, 1\}^n$, and we can simplify the energy function from Eq. (11) to Eq. (14).

$$E(\boldsymbol{l}) = \sum_i \phi_i'(l_i) + \beta \sum_{i<j} \psi_{i,j}'(l_i, l_j), \tag{14}$$

$$\phi_i'(l_i) = \begin{cases} \phi_i'(1) = \sum_p \phi_i^p(1) = \sum_p \varsigma_\alpha(-d_i^p) \\ \phi_i'(0) = \sum_p \phi_i^p(0) = \sum_p (1 - \varsigma_\alpha(-d_i^p)), \end{cases} \tag{15}$$

$$\psi_{i,j}'(l_i, l_j) = \sum_p \psi_{i,j}(l_i, l_j) = m\psi_{i,j}(l_i, l_j). \tag{16}$$

The global minimum of Eq. (14) can be obtained by using QPBO [17].

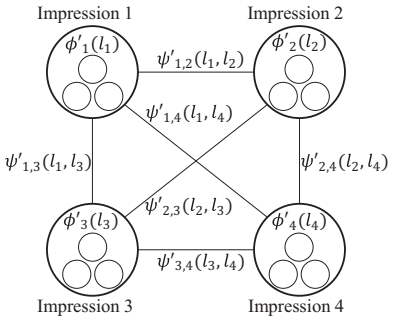


Fig. 4: The graph structure of Eq. (14). All layers in Fig. 3 have degenerated into one layer.

According to the deformation of the energy function from Eq. (11) to Eq. (14), the graph in Fig. 3 degenerates into the one in Fig. 4; in other words, the m layers in Fig. 3 degenerate into one layer in Fig. 4. This graph has n nodes, and each node includes m subnodes. The p-th subnode in the i-th node in Fig. 4 corresponds to the i-th node in the p-th layer in Fig. 3. The label l_i is assigned to the i-th node. The i-th node has the unary cost $\phi'_i(l_i)$ in Eq. (15). The edge connecting the i-th and the j-th nodes corresponds to the pairwise cost $\psi'_{i,j}(l_i, l_j)$ in Eq. (16) and represents label relations between impressions. In the following experiments about late feature fusion in Sec. 4, we use the energy function in Eqs. (14)-(16).

4. Experimental Results

Although the proposed method is general and also can be employed in other applications of multi-label classification problems, here we conduct the experiments that deal with the impression prediction for TED Talks. The impression analysis of oral presentations based on linguistic and acoustic features has been studied in our laboratory [25]. In this experiment, we apply the proposed method to this impression prediction of oral presentations [25], and confirm the usefulness of the proposed method.

4.1 Dataset

There are more than 1,900 videos in TED Talk. We eliminated non-oral-presentation types of talks such as playing music, magic shows, showing visual content such as cartoons, and so on. As a result 1,646 presentations were used in the experiments. Viewers on the Internet can vote for three impressions out of the 14 types of impressions: *beautiful*, *confusing*, *courageous*, *fascinating*, *funny*, *informative*, *ingenious*, *inspiring*, *jaw-dropping*, *longwinded*, *obnoxious*, *OK*, *persuasive*, and *unconvincing*. If and only if the viewer votes only for one impression, it is counted as three votes. All the presentation videos, their transcripts, and the impression rates were downloaded by using the API.

4.2 Features

We use three types of features for training and testing of SVM.

Content Features: Bag-of-Words (BoW) [14] representation of texts is one of the simplest but efficient text representations, where the frequency of each tag is counted to form a histogram. Latent Semantic Indexing (LSI) [9] and Latent Dirichlet Allocation (LDA) [6] are dimensionality reduction techniques assuming that the words that are used in similar contexts tend to have similar meanings. We extract BoW features, latent dimensional representations from LSI, and latent topic representations from LDA as features from the transcripts of oral presentations.

Motivated by the impressive success of word representations in related classification tasks such as sentiment classification [22], we also use the skip-gram method implemented in word2vec[*2] (w2v) tool to learn representations for words.

Table 1: Details of the surface-level text features.

Feature	Dim.	Description
Ave. # of words in a sentence	1	
Ave. # of characters in a word	1	
Ave. # of syllables in a word	1	
Total # of sentences	1	
Total # of words	1	
Total # of characters	1	
Total # of syllables	1	
Hist. of # of words in sentences	12	1-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-18, 19-20, 20-21, 23-28, 29-40, 41-
Hist. of # of characters in words	11	1, 2, ..., 10, 11-
Hist. of # of syllables in words	5	1, 2, 3, 4, 5-
In which school year the words are learned	9	1st year, 2nd year, .., 8th year, SAT
Total	44	

Surface-level linguistic features: In addition to the content of the talk,how dignifiedly the speaker talks is also an important factor to analyze the impression. The features we defined are listed in Table 1. We used the school vocabulary list provided in BigIQkids [1] to find at which year of school does a student learns a particular word for the first time and use this information as a feature that encodes the language fluency of a speaker.

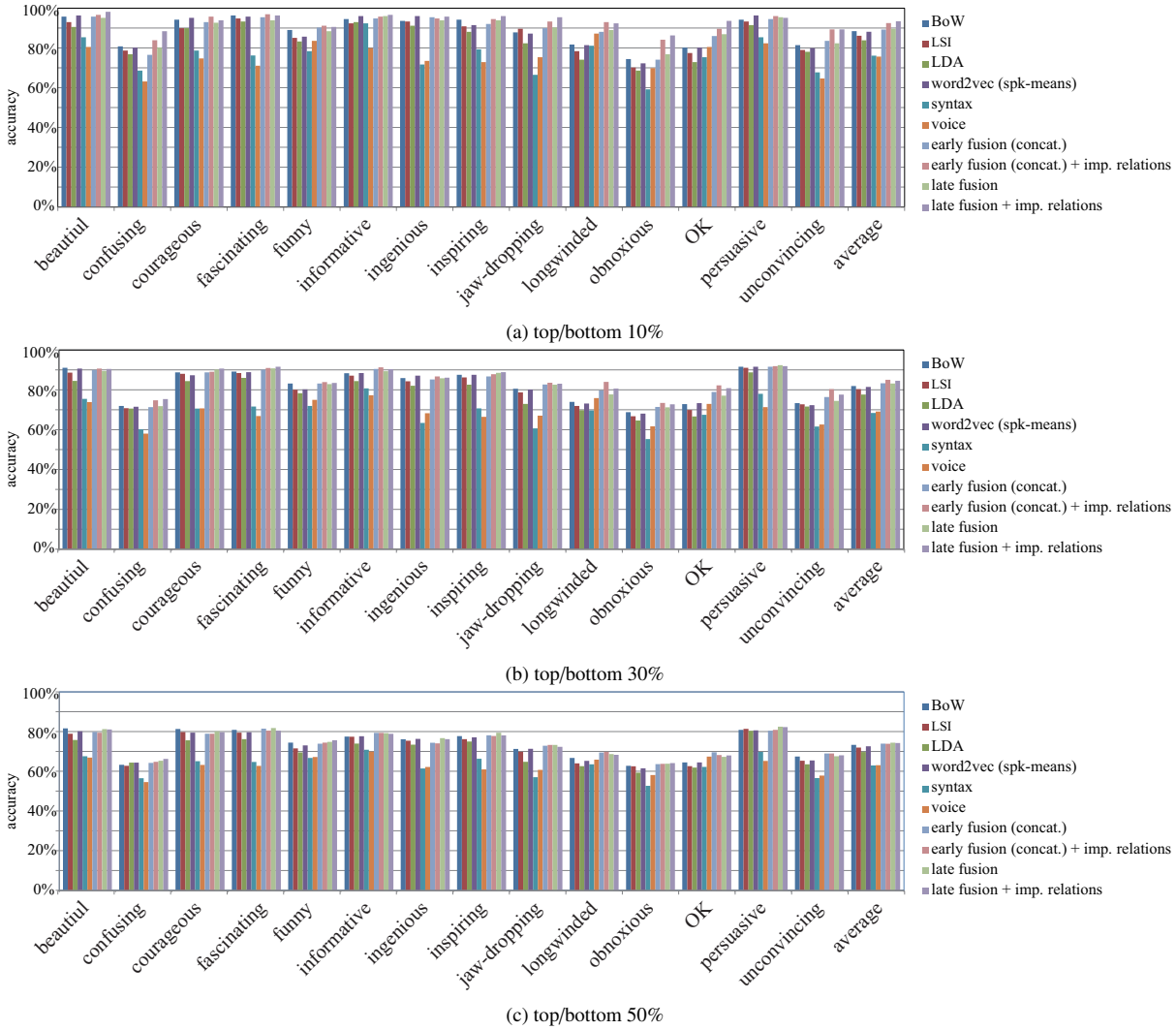
Acoustic Features: The impression of the presentation is also determined by how the speaker talks. We employed openSMILE [12] to extract acoustic features. The configuration was the same as that in INTERSPEECH 2013 Computational Paralinguistics Challenge[*3]. As a result, a 6,373-dimensional feature vectors including pitch, voice quality, energy, loudness, spectral, MFCC, etc. are extracted from each presentation.

In the experiments, we use all articles from an English Wikipedia snapshot collected in 2015 to train LSI, LDA, and skip-gram, and use the top 100,000 frequently used words for representing the articles. The dimension of the feature vector for LSI and LDA was swept from 100 to 3,000, respectively, and from 100 to 30,000 for skip-gram. α in Eq. (2) was set to 3.0 throughout the experiments.

4.3 Classification Results

Figure 5 shows the accuracy of the impression prediction. The accuracy was calculated by the leave-one-out method using the SVM with a radial basis function (RBF) kernel. The parameters for the SVM were optimized by the grid search in advance. Considering that impression rates are continuous values, it is often

[*2] https://code.google.com/p/word2vec
[*3] http://emotion-research.net/sigs/speech-sig/is13-compare



(a) top/bottom 10%

(b) top/bottom 30%

(c) top/bottom 50%

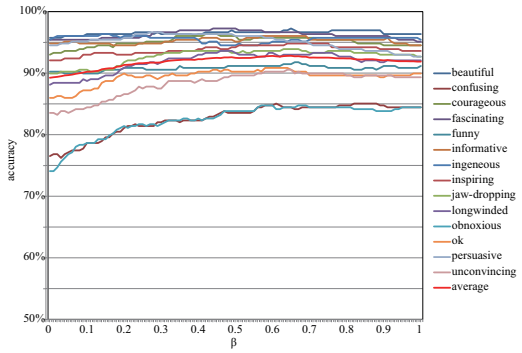
Fig. 5: Impression prediction accuracy, where $\beta = 0.3$.

difficult to determine a threshold value to binarize the continuous impression rates to positive vs. negative classes. Therefore, instead of evaluating using an arbitrary single threshold point, we select the top and bottom r% of the videos according to their impression rates respectively as the positive and negative instances. As shown in Fig. 5, we conduct this analysis for $r = 10, 30$, and 50.

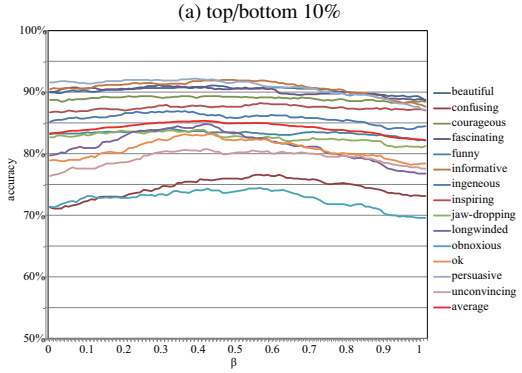
In Fig. 5(a), only the top and bottom 10% of the TED Talk videos for each impression class were used in the experiment (i.e., 165 videos with high ratio of a certain impression and 165 with low ratio, 330 videos in total.). In Fig. 5(b), the top and bottom 30% were used. In Fig. 5(c), all the videos were used and the top half was labeled as "positive" and the bottom half as "negative". As described in Sec. 4.2, the dimension for LSI and LDA was changed from 100 to 3,000 and that for skip-gram from 100 to 30,000 according to the previous approaches and only the best dimensions in terms of average accuracy are employed (3,000 for LSI and LDA, and 10,000 for skip-gram). It is shown that the content-based linguistic features (i.e., BoW, LSI, LDA, and skip-gram) generally outperform surface-level linguistic features and acoustic features when the features are used independently. There is only a slight improvement in the early feature fusion where the features are simply concatenated. A large improvement can be observed in our proposed method. On average, the accuracy is improved from 89.2% to 93.3%, from 83.3% to 84.5%, and from 73.9% to 74.2% in the top and bottom 10%, 30%, and 50% cases, respectively, when both correlations between different impression labels, and correlations between different feature types are considered. The best performance is achieved when only the correlations between different impression labels is considered for the case of $r = 30$ (85.0%), and when only the correlations between different feature types is considered for the case of $r = 50$ (74.4%).

It is also shown that the accuracy of the impression prediction improves for most of the impression types as compared to the baseline methods. As the top-and-bottom ratio, r, increases, the improvements of the prediction accuracy become smaller. This is a matter of course because the presentations near the positive and negative boaders are similar to each other in terms of the vote rates.

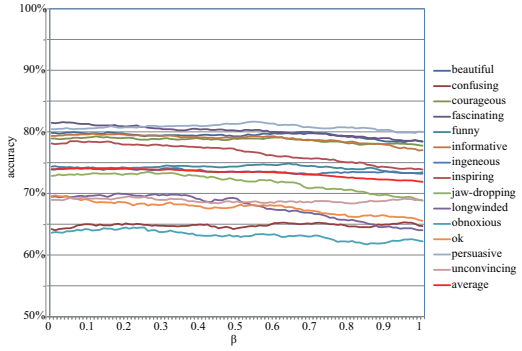
The performance of the impression prediction is shown as a function of β in Fig. 6. $\beta = 0$ corresponds to the early fusion. It can be observed that the prediction is improved as the β is increased up to a certain point ($\beta = 0.3$) and gradually get degraded because the label relationships become more dominant than the label outputs from the classifiers.



(a) top/bottom 10%

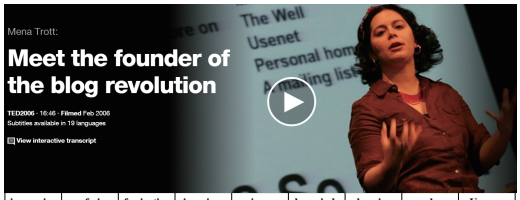


(b) top/bottom 30%



(c) top/bottom 50%

Fig. 6: Impression prediction accuracy as a function of β.



impression	confusing	fascinating	ingenious	jaw-dropping	longwinded	obnoxious	ok	Unconv.
label	NO	NO	NO	NO	YES	YES	NO	NO
dec. value	-0.088	-0.610	-0.718	-0.728	0.451	0.481	-0.116	-0.190
incorrect	incorrect	correct	correct	correct	correct	correct	incorrect	incorrect
⇩								
re-label	YES	NO	NO	NO	YES	YES	YES	YES
incorrect	correct	correct	correct	correct	correct	correct	correct	correct

(a) Talk ID: 21

impression	beautiful	persuasive
label	NO	NO
dec. value	-0.696	-0.057
correct or incorrect	correct	incorrect
⇩		
re-label	NO	YES
correct or incorrect	correct	correct

(b) Talk ID: 193

impression	courageous	fascinating	jaw-dropping	ok	unconvincing
label	NO	YES	NO	NO	YES
dec. value	-0.855	0.932	-0.068	-0.129	0.246
correct or incorrect	correct	correct	incorrect	correct	incorrect
⇩					
re-label	NO	YES	YES	NO	NO
correct or incorrect	correct	correct	correct	correct	correct

(c) Talk ID: 509

Fig. 7: Examples of successful cases.

Fig. 7 shows some of the successful cases. In Fig. 7(a), negative impressions are successfully relabeled as "yes" by considering the label relations. As can be observed in Fig. 5, negative impressions are more difficult to predict than positive impressions. By our proposed method, the correlation among the impressions are successfully contributing to updating the impression labels. We found that there is a strong negative correlation between *beautiful* and *pursuasive* (the correlation coefficient was −0.41.) Therefore, our model updated the impression label from *not pursuasive* to *pursuasive* in Fig. 7(b). In Fig. 7(c), both positive and negative impressions are updated to correct ones by considering the other impression labels.

5. Conclusions

In this paper, a joint optimization framework for multi-label classification and late feature fusion based an MRF-based formulation has been proposed and successfully applied to impression prediction for TED Talks videos by combining linguistic and acoustic features. In our proposed method, the label relationships were softly incorporated into the pair-wise term. The late feature fusion has been achieved by adding "must-co-occur" constraint between classifiers. For the impression prediction of oral presentations, state-of-the-art linguistic and acoustic features were effectively and efficiently combined. The impression prediction accuracy of 93.3% has been achieved for the top/bottom 10% data. Although the multi-label and late feature fusion techniques were applied to TED Talks analysis, we believe that the algorithms are general and applicable to a lot of different applications.

In our future work, we are planning to apply the proposed method to other applications of multi-label classification problems, and confirm the generality of our method. We are also planning to extend the proposed method to the regression problem, where the confidence scores of assigned labeling are estimated. In addition, we will apply our MRF formulation to multi-class object recognition. Our method will be able to take account of the relationships between different classes successfully.

References

[1] BigIQkids. `http://www.bigiqkids.com/`.

[2] TED Talks. https://www.ted.com/.
[3] M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundat. TrendsMach. Learn.*, 4(3):195–266, 2012.
[4] W. Bi and J. T. Kwok. Multi-label classification on tree-and dag-structured hierarchies. In *ICML*, pages 17–24, 2011.
[5] W. Bi and J. T. Kwok. Mandatory leaf node prediction in hierarchical multilabel classification. In *NIPS*, pages 153–161, 2012.
[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
[7] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.
[8] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, and T.-S. Chua. Multi-label visual classification with label exclusive context. In *ICCV*, pages 834–841, 2011.
[9] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
[10] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.
[11] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*, pages 48–64. 2014.
[12] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACMMM*, pages 835–838, 2013.
[13] M. Gonen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.
[14] Z. S. Harris. Distributional structure. *Word*, 10:146–162, 1954.
[15] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE TMM*, 12(6):523–535, Oct 2010.
[16] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, pages 1719–1726, 2006.
[17] V. Kolmogorov and C. Rother. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE TPAMI*, 29(7):1274–1279, 2007.
[18] D. Liu, K.-T. Lai, G. Ye, M.-S. Chen, and S.-F. Chang. Sample-specific late fusion for visual category recognition. In *CVPR*, pages 803–810, 2013.
[19] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084 – 3104, 2012.
[20] K. Nandakumar, Y. Chen, S. Dass, and A. Jain. Likelihood ratio-based biometric score fusion. *IEEE TPAMI*, 30(2):342–347, 2008.
[21] T. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.
[22] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP*, pages 151–161, 2011.
[23] O. Terrades, E. Valveny, and S. Tabbone. Optimal classifier fusion in a non-bayesian probabilistic framework. *IEEE TPAMI*, 31(9):1630–1644, 2009.
[24] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*, pages 721–728, 2002.
[25] T. Yamasaki, Y. Fukushima, J. Xu, and S. Sakazawa. Impression estimation of oral presentations based on document and voice analysis. In *MVE-2014-99*, pages 119–122. IEICE MVE, 2015. (In Japanese).
[26] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE TKDE*, 18(10):1338–1351, 2006.
[27] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038 – 2048, 2007.