CONSTRAINED MINIMIZATION TECHNIQUE FOR TOPIC IDENTIFICATION USING DISCRIMINATIVE TRAINING AND SUPPORT VECTOR MACHINES

Imed Zitouni∗, Minkyu Lee2, Hui Jiang3

1 IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
2 Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA
3 York University, Toronto, Ont. M3J 1P3, CANADA

izitouni@us.ibm.com, minkyul@lucent.com, hj@cs.yorku.ca

ABSTRACT

This paper describes the constrained minimization approach to combine multiple classifiers in order to improve classification accuracy. Since errors of individual classifiers in the ensemble should somehow be uncorrelated to yield higher classification accuracy, we propose a combination strategy where the combined classifier accuracy is a function of the correlation between classification errors of the individual classifiers. To obtain powerful single classifiers, different techniques are investigated including support vector machines and latent semantic indexing (LSI) matrix, which is a popular vector-space model. We also investigate discriminative training (DT) of the LSI matrix on constrained minimization approach. DT minimizes the classification error by increasing the score separation of the correct from competing documents. Experimental evaluation is carried out on a banking call routing and on switchboard databases with a set of 23 and 67 topics respectively. Results show that the combined classifier we propose outperforms the accuracy of individual baseline classifiers by 44%.

1. INTRODUCTION

Automatic topic identification is a classification problem that assigns a topic label to a document. The document may be either a text stream or a spoken message. In the latter case, we can apply a text-based classifier by passing the spoken message through a speech recognizer first. One application to automatic topic identification is to automatically identify the type of a call requested by a customer. Indeed, call centers are currently looking for natural language call routing applications, where the caller may say what he/she wants and is automatically routed to the right department or directed to a human operator when the system is unable to determine the caller’s intent with certainty.

Several papers have been published on automatic topic identification [1, 2, 3]. Recently, we presented new techniques such as discriminative training (DT) on the vector-based model to improve classification accuracy and robustness of single classifiers [3, 2]. In this paper, we investigate the combination of multiple classifiers to achieve better performance than any individual classifier. We consider the constrained minimization technique on singular classifiers where their errors are uncorrelated. This technique was introduced on [7] using artificial data and binary classifier for nasal/oral vowel classification task. We investigate here its effectiveness on real application with multi-class task and real data. To make errors of single classifiers uncorrelated, one approach consists of re-sampling the data with different distribution and training separate classifiers on each re-sampled data set [3]. To start with accurate single classifiers, we use support vector machines (SVMs) and DT on the vector-based model.

At the outset, it is clear that multiple copies of the same classifier are no better than one classifier. Therefore the classifiers must clearly be different from each other in some respect. The important questions are: (i) how should the classifiers be different from each other and how can we engineer them to be so, (ii) how should they then be combined, and (iii) how does the error rate after combination relate to the achievable error rates with individual single classifiers. To keep matters simple and tractable, we consider a three-classifier system and analyze the possibilities contained therein. Experimental evaluation is carried out on a banking call routing task and on switchboard databases containing 23 and 67 topics respectively.

2. BASELINE VECTOR-BASED CLASSIFIERS

2.1. Cosine Classifier

The classifier using the cosine similarity metric is a popular vector-based classifier [1]. The training process involves constructing a routing matrix \( R (m \times n) \). The rows of \( R \) represent the \( m \) terms (e.g., words) and the columns \( (r_j^T) \) the \( n \) destinations. The routing matrix \( R \) is the transpose of the term-document matrix, where each element \( r_{vw} \) represents the frequency of term \( w \) that occurs in calls to destination \( v \). Each term is weighted according to the term frequency inverse document frequency (TFIDF) and is also normalized to unit length. Documents are represented as feature vectors and are labeled based on the cosine similarity score. Let \( \tilde{x} \) be the \( m \)-dimensional observation vector representing the weighted terms which have been extracted from the document. One possible labeling decision is to attribute the topic that has the highest cosine similarity score:

\[ \text{destination } j = \arg \max_j \cos \phi_j = \arg \max_j \frac{r_j^T \cdot \tilde{x}}{\|r_j\| \|\tilde{x}\|} \]  

(1)

2.2. Beta classifier

The beta classifier is a probabilistic method, which has previously been shown to give the best results in a study on e-mail routing [4]. Each topic is represented by a word lexicon. For each word in the lexicon we compute its probability in the topic and its weight [4]. This weight is assigned according to a function inversely proportional to the number of topic-lexicons in which this word is.

∗This work was done when Dr. Zitouni was a Reasearch Member of Technical Staff at Bell-Labs
where \( P(w_k|j) \) is the probability of \( w_k \) in topic \( j \), and \( \eta(w_k) \) the weight assigned to \( w_k \). Parameters \( \delta_1 \) and \( \delta_2 \) are estimated on a development corpus to boost the accuracy. In our experiments, we obtain a value of 0.3 for \( \delta_1 \), a value of 2 for \( \delta_2 \) and we take into account words that occur at least three times in the corpus. The term \( \beta_j \) is the weight assigned to topic \( T_j \):

\[
\beta_j = \frac{\sum_{k=1}^{N_k} \eta(w_k)}{\sum_{k=1}^{N_k} \eta(w_k)}.
\]

where \( N_k \) represents the number of words in the \( k \)th topic-vocabulary.

3. SUPPORT VECTOR MACHINE CLASSIFIERS

Support vector machines (SVMs) have already been used for text categorization, where they showed to achieve substantial improvement over state-of-the-art methods [5]. SVMs are based on the Structural Risk Minimization principle. As such, it is firmly grounded in the framework of statistical learning theory, or Vapnik-Chervonenkis (VC) theory [6]. The idea of structural risk minimization is to find a hypothesis \( h \) for which we can guarantee the lowest true error \( \epsilon \). The true error of \( h \) is the probability that \( h \) will make an error on a randomly selected unseen test document. SVMs find the hypothesis \( h \) which minimizes a bound on the true error by efficiently and effectively controlling the VC dimension of the hypothesis space [6].

4. DISCRIMINATIVE TRAINING TECHNIQUE

DT has recently been proposed for natural language call routing [2] and has been shown to be highly effective in simplifying the classifier design and improving portability [2, 3]. Instead of simply counting in conventional maximum likelihood training, the minimum classification error criterion is used in DT of the routing matrix parameters. Classification accuracy and robustness are improved by adjusting the models to increase the separation of the correct class from its competitors. The same framework is used in this paper.

5. CONSTRAINED MINIMIZATION TECHNIQUE

The possibility of building multiple classifiers and then combining them to obtain a more accurate one is of considerable interest. In this paper, we consider a combination strategy for three different classifiers [7] whose errors are supposed to be uncorrelated. Suppose we have two classifiers \( C_1 \) and \( C_2 \) which predict the topics \( t_1 \) and \( t_2 \) respectively for a document or a query \( q \). When both classifiers agree (\( t_1 = t_2 \)), the topic result is the same as each of the classifiers. When they disagree, a third classifier is invoked as an arbiter. This third classifier may be explicitly trained on disagreements of the first two using minimum error training or it can also make a choice only on a subset of topics. This subset may be computed according to the N-best topics proposed by each of the first two classifiers or according to a confusion measure. For example, when the first two classifiers disagree, we take the set of confusable topics \( ST_1 = \{ t_{i_1} \} \) in \( C_1 \) and \( ST_2 = \{ t_{i_2} \} \) in \( C_2 \). Then, the third classifier \( C_3 \) chooses among the topics in the union of these subsets \( (ST_1 \cup ST_2) \). Let \( t_{C_i} \) denote the best topic chosen by the classifier \( C_i (i \in \{1, 2\}) \) to the document \( q \). Hence, the set of confusable topics for \( C_i \) are those \( ST_i = \{ t_{i} \} \) with a distance to \( t_{C_j} \) smaller than a threshold \( \theta \) computed according to the average distance between \( t_{C_i} \) and the correct one \( t^*_{C_i} \) on the training set:

\[
d(l_{C_i}, t^*_{C_i}|q) < \theta.
\]

In this paper we choose the Kullbach-Leiber distance:

\[
d(t_1, t_2|q) = S_{C_i}(t_1, q) \ln \left( \frac{S_{C_i}(t_2, q)}{S_{C_i}(t_1, q)} \right),
\]

where \( S_{C_i}(t, q) \) denotes the probability (score) attributed to the topic \( t \) for the document \( q \) with the classifier \( C_i \).

We will next discuss the performance of the combined classifier using this approach, according to the accuracy of each classifier \( (C_1, C_2, C_3) \) and the correlation between them. Without loss of generality, we consider in the following the binary classification problem. Let \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) be the data set, where \( x \) is the document and \( y \) is its topic. Let \( e_{C_i} \) be the error function associated with the classifier \( C_i \); defined as a boolean function. It takes the value 1 if the classifier makes an error and 0 otherwise:

\[
e_{C_i}(x, y) = \begin{cases} 1 & \text{if } C_i(x) \neq y, \\ 0 & \text{otherwise}. \end{cases}
\]

The error rate \( \epsilon_i \) associated with classifier \( C_i \) is defined by the mathematical expectation on \( e_{C_i} : \epsilon_i = E(e_{C_i}) \). The correlation between the two classifiers \( C_1 \) and \( C_2 \) can be represented in terms of the covariance \( \Psi(C_1, C_2) \) between their respective errors:

\[
\Psi(C_1, C_2) = E\left((e_{C_1} - E(e_{C_1}))(e_{C_2} - E(e_{C_2}))\right) = \rho - \epsilon_1\epsilon_2
\]

where \( \rho \) denotes the probability that both classifiers \( C_1 \) and \( C_2 \) are wrong:

\[
\rho = E(e_{C_1} e_{C_2}).
\]

The error probability \( \Upsilon \) of the combined classifier is the sum of:

- the probability that both \( C_1 \) and \( C_2 \) are wrong: \( \rho \),
- the probability that they disagree and the third classifier \( C_3 \) is wrong:

\[
\left( (\epsilon_1 - \rho) + (\epsilon_2 - \rho) \right) \epsilon_3.
\]

Adding the two and putting in \( \Psi \) (equation (7)), the error rate of the combined classifier \( \Upsilon \) becomes as follows:

\[
\Upsilon = (\epsilon_1\epsilon_2 + \Psi) (1 - 2\epsilon_3) + (\epsilon_1 + \epsilon_2) \epsilon_3.
\]

The probability \( \rho \) is positive and smaller than \( \epsilon_1 \) as well as \( \epsilon_2 \) \( (\rho \leq \epsilon_1 \) and \( \rho \leq \epsilon_2) \). Therefore, the correlation between classifier \( C_1 \) and \( C_2 \) is bounded by [7]:

\[
-\epsilon_1\epsilon_2 \leq \Upsilon \leq \min(\epsilon_1, \epsilon_2) - \epsilon_1\epsilon_2.
\]

From equations 9 and 10, the following bound on the error rate of the combined classifier \( \Upsilon \) is obtained:

\[
(\epsilon_1 + \epsilon_2)\epsilon_3 \leq \Upsilon \leq \min(\epsilon_1, \epsilon_2) + |\epsilon_2 - \epsilon_1|\epsilon_3
\]

where \(|.|\) denotes the absolute value function.

It is important to note that \( \Upsilon \) is not guaranteed to be less than the minimum of the individual classifiers [7]. Hence, we conclude that:

\[
\text{...}
\]
• the overall error rate of the combined classifier depends upon the correlation between them and the strengths of the individual classifiers;

• improvement is guaranteed only if their errors are sufficiently uncorrelated;

• the correlation between the classifiers can be captured by the covariance \( \Psi \) which is related to \( \rho, \varepsilon_1, \) and \( \varepsilon_2. \)

6. EXPERIMENTS

Experiments were performed on two topic identification tasks, a DARPA Switchboard text categorization task and a banking call routing task with USAA. We use the switchboard database which was initially released, consisting of a total of 2284 transcribed conversations and 67 topics [8]. We divided the database into a training set consisting of about 80% of the database, and the remaining 20% was used as the test set. The lexicon used contains 5000 most frequent words. For the banking call routing task, we used the same training and test sets as reported in [1], consisting of a total of about 4000 calls, routed to 23 destinations. The vocabulary used for the banking task contains 1232 words; each one of them appears more than three times in the training data. Experimental results on the banking task are reported on both human transcriptions (Banking-HT) and ASR recognized strings (Banking-ASR). We used real-time speech recognition system, which have a word error rate of about 20% [3]. Some results are not the same as previously reported in [2] because a different set of unigram features is used in this paper. Results on switchboard database are not comparable with published results for many reasons, one of which is that previous experiments used only 10 topics [9]. SVM experiments were conducted using software from Royal Holloway and Bedford New College, University of London [10]. We used the 1-vs-1 approach to multi-class classification and we limit ourselves to the use of dot product in \( \mathbb{R}^n \) as a kernel [11].

6.1. Performance of Individual Classifiers

We want to investigate how much improvement can be achieved using DT on classifiers that contain only simple features, consisting of the most common words. We also show a comparison study between DT and SVMs. Table 1 shows the classification error rate (CER) of SVMs, beta and cosine classifiers with and without DT. Results indicate that DT improves classifier accuracy of the baseline classifiers (cosine and beta) by 35% on the banking task. For the errorful strings obtained from speech recognition, the relative error rate reduction is about 30%. On switchboard task, we obtain an average of 69% relative improvement when DT is used. Results show that DT is effective even in cases where tasks or data to be processed (i.e., errorful strings from ASR system) are different.

Results also show that DT applied on vector-based model outperforms SVMs by 6% on average. Because of the relatively little training data in both tasks, we cannot confirm the hypotheses that DT works better than SVMs. However, we believe that these results again testify to the effectiveness of DT to achieve good performance. Zitouni et al. in [3] showed how DT can obtain better performance compared to other approaches, including boosting and relevance feedback. In this paper, results confirm once again that DT is able to achieve very competitive accuracy compared to other techniques, including SVMs.

<table>
<thead>
<tr>
<th></th>
<th>Switchboard</th>
<th>Banking-HT</th>
<th>Banking-ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>19.1%</td>
<td>9.4%</td>
<td>12.0%</td>
</tr>
<tr>
<td>Cosine+DT</td>
<td>5.9%</td>
<td>6.1%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Beta</td>
<td>18.0%</td>
<td>12.0%</td>
<td>14.9%</td>
</tr>
<tr>
<td>Beta+DT</td>
<td>5.9%</td>
<td>5.5%</td>
<td>7.8%</td>
</tr>
<tr>
<td>SVMs</td>
<td>6.3%</td>
<td>6.5%</td>
<td>8.4%</td>
</tr>
</tbody>
</table>

Table 1. Classification error rate of SVMs, cosine, and beta classifiers with and without DT.

6.2. Multiple Classifier Combination

In the following, we investigate different methods to combine the classifiers we have: SVMs, beta, and cosine classifiers (with and without DT). First, linear interpolation (LI) of the different classifiers is investigated; we use a linear combination of individual classifier scores. Then, the constrained minimization technique is employed to combine these classifiers.

6.2.1. Constrained Minimization Technique

The constrained minimization technique is investigated to combine beta and cosine classifiers: we consider \( C_1 \) and \( C_2 \) as classifiers based on beta and cosine metrics respectively. When both classifiers agree, the topic result is the one agreed upon. When they disagree, a third classifier \( C_3 \) is invoked as an arbitrator. The third classifier \( C_3 \) is usually trained on the disagreements of the first two [7]; the classifier type for \( C_3 \) is chosen to be the one with the smaller classification error rate, which is cosine in our case. This classifier \( C_3 \) proceeds only on a subset of topics chosen according to a confusion measure (cf. §5); we denote this approach \( CM_D \). The third classifier \( C_3 \) of \( CM_D \) will choose among a topic set, which is the union of confusable topics of each individual classifier. The confusable topics of a classifier \( C_i \) are those with a distance to the best topic (predicted by \( C_i \)) smaller than a threshold. Another experiment is also done in which \( C_3 \) makes a choice among the N-best topics from \( C_1 \) and \( C_2 \); denoted \( CM_N \). In our experiment, the value of \( N \) is set to 2 which gives the best result.

<table>
<thead>
<tr>
<th></th>
<th>LI</th>
<th>CM_D</th>
<th>CM_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine + beta</td>
<td>17.8%</td>
<td>16.9%</td>
<td>17.1%</td>
</tr>
<tr>
<td>(cosine+beta) w' DT</td>
<td>5.9%</td>
<td>5.7%</td>
<td>5.7%</td>
</tr>
<tr>
<td>Banking Human Transcription</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine + beta</td>
<td>10.4%</td>
<td>7.8%</td>
<td>8.1%</td>
</tr>
<tr>
<td>(cosine+beta) w' DT</td>
<td>5.8%</td>
<td>5.8%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Banking ASR Recognized Strings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine + beta</td>
<td>12.7%</td>
<td>10.0%</td>
<td>10.4%</td>
</tr>
<tr>
<td>(cosine+beta) w' DT</td>
<td>7.8%</td>
<td>8.1%</td>
<td>8.1%</td>
</tr>
</tbody>
</table>

Table 2. CER of different combination techniques between cosine and beta, with and without the use of DT.

We present in Table 2 the classification error rate of the combined classifiers. Experiments show that the combination between these two classifiers is a good way to improve the performance. On the banking human transcription data, the constrained minimization technique with beta and cosine classifiers (7.8%) increases the accuracy of the best single classifier (9.4%) by 17%. An improvement of 6% is also achieved on the switchboard task when
the constrained minimization technique is used (18% vs. 16.9%). However, almost no improvement is reported when the linear interpolation is used between beta and cosine classifiers. It is important to note that when we apply DT on the baseline classifiers, the constrained minimization technique does not seem to be able to reach further improvement. Results show that the use of a confusion measure technique works slightly better than the fact to make the third classifier chose between the N-best topics of each of the first two.

### 6.2.2. Improvement of Constrained Minimization Technique

Constrained minimization did not give the improvement we had expected when DT is employed. The reason is that the CER of the third classifier $C_3$ (trained on the disagreement set) is quite high on the entire test set, about 65%. Hence, to better use this technique, we build a new classifier $C_3$ with a higher accuracy. Let classifiers $C_1$ and $C_2$ represent SVMs and cosine classifier discriminatively trained on the entire training corpus, respectively. Then, let $C_3$ represent the beta classifier with and without DT. One motivation of using SVMs for $C_2$ is the fact that its errors are not correlated with the classifiers discriminatively trained. Let $CM^*$ denote the combined classifier using beta classifier as $C_3$, and let $CM^*_DT$ denote the combined classifier using beta classifier discriminatively trained as $C_3$. We use the confusion measure technique for $C_3$, which gave better results (cf. §5). As a reminder, the third classifier $C_3$ will choose among a topic set, which is the union of confusable topics of each individual classifier.

<table>
<thead>
<tr>
<th></th>
<th>$CM^*$</th>
<th>$CM^*_DT$</th>
<th>LI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Switchboard</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Classifier</td>
<td>5.4%</td>
<td>5.0%</td>
<td>5.9%</td>
</tr>
<tr>
<td>(SVM + Cosine w' DT)+Beta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Banking Human Transcription</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Classifier</td>
<td>5.5%</td>
<td>5.2%</td>
<td>5.8%</td>
</tr>
<tr>
<td>(SVM + Cosine w' DT)+Beta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Banking ASR Recognized Strings</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Classifier</td>
<td>7.4%</td>
<td>7.1%</td>
<td>7.8%</td>
</tr>
<tr>
<td>(SVM + Cosine w' DT)+Beta</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. CER of three classifiers using linear interpolation (LI) and constraint minimization ($CM^*$ and $CM^*_DT$).

Table 3 presents the CER of this combination as well as a linear interpolation between these three classifiers ($C_1, C_2, C_3$). As expected, the accuracy of $CM^*$ and $CM^*_DT$ is better than those cited before in Table 2 and is also better than linear interpolation between the different classifiers. Experiments show that DT again boost the classification accuracy: $CM^*_DT$ outperforms $CM^*$ by an average of 8%. On the switchboard database, the combination of multiple classifiers $CM^*_DT$ (5.0%) improves the baseline version discriminatively trained (5.9%) by 15%. For the errorful strings obtained from speech recognition on the banking task, $CM^*_DT$ (7.1%) outperforms the baseline version discriminatively trained (7.8%) by 9%. A comparative improvement is also obtained using the human transcribed banking data. Other experiments are also done using constrained minimization technique, were we changed the position of the three classifiers $C_1, C_2$ and $C_3$. Results are quite similar. The use of constrained minimization with DT and SVMs showed considerable improvement of the classification accuracy. On the banking task, compared to cosine and beta classifiers, with CER of 9.4% and 12% respectively, more than 44% improvement is obtained when we use $CM^*_DT$ (5.2%).

### 7. Conclusion

We have investigated in this paper a combination strategy on individual classifier using the constrained minimization approach. We showed how the overall error rate of the combined classifier depends upon the correlation between individual classifier errors and their strengths. An improvement of the combined classifier accuracy is guaranteed only if their errors are sufficiently uncorrelated. One idea of making errors of single classifiers uncorrelated consists of re-sampling the data with different distribution and training separate classifiers on each re-sampled data set. To obtain state-of-the-art individual classifiers, SVMs and DT on vector-based model are used. Results show clearly how DT is able to achieve very competitive results compared to other approaches, including SVMs. Experiments on the banking and switchboard tasks show that the approach we propose, which includes SVMs and DT, outperforms the accuracy of baseline classifiers by 44%. A 15% improvement is also obtained compared to baseline classifiers discriminatively trained.

### 8. References


