ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Constrained Maximum Mutual Information Dimensionality Reduction for Language Identification

Shuai Huang[1], *Glen A. Coppersmith*[1,2], *Damianos Karakos*[3]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA
[3]Raytheon BBN Technologies*, Cambridge, MA, USA
{shuaihuang, coppersmith}@jhu.edu, dkarakos@bbn.com

10.21437/Interspeech.2012-544

Abstract

In this paper we propose Constrained Maximum Mutual Information dimensionality reduction (CMMI), an information-theoretic based dimensionality reduction technique. CMMI tries to maximize the mutual information between the class labels and the projected (lower dimensional) features, optimized via gradient ascent. Supervised and semi-supervised CMMI are introduced and compared with a state of the art dimensionality reduction technique (Minimum/Maximum Rényi's Mutual Information using the Stochastic Information Gradient; MRMI-SIG) for a language identification (LID) task using CallFriend corpus, with favorable results. CMMI also deals with higher dimensional data more gracefully than MRMI-SIG, permitting application to datasets for which MRMI-SIG is computationally prohibitive.

1. Introduction

The prevalence of inference tasks which must operate on high dimensional data is increasing with no sign of abatement. Although it is usually beneficial to have more information with which to tackle such tasks, the "curse of dimensionality" brings with this added number of dimensions concerns about data sparsity and raises the computational requirements. Dimensionality reduction is often employed to address these concerns by projecting the data onto a lower dimensional space, while preserving useful hidden structures in the original data.

In [1] a supervised approach that maximizes the mutual information between the target variable (class label) C and the projected data P^TX: $I(C;P^TX)$ was proposed. It tries to find orthogonal normalized 1-dimensional projections one by one. This iterative projection approach is a greedy method and the mutual information is approximated through histogram estimation, which might pose another computational problem between accuracy and efficiency. Similarly, Minimum/Maximum Rényi Mutual Information using the Stochastic Information Gradient (MRMI-SIG) [2] is another supervised approach which yields state of the art classification performance for dimensionality reduction. However, it is not designed to handle very high-dimensional dataset and suffers similar problems as [1].

Consider a dataset X ($|X| = N$ documents) with high dimensional features W ($|W| = M$), with class labels C. We propose the Constrained Maximum Mutual Information dimensionality reduction (CMMI), which attempts to find a transformation matrix P that maximizes the mutual information between the class labels C and a set of features projected into a lower dimension W: $I(C;W')$ where $W' = P^TW$. Note that, as opposed to [2], no discretization of the data using Parzen windows is needed; each data point is a discrete distribution.

The paper proceeds as follows: a proof of the convexity of the objective function $I(C;W')$ is given in Section 2, followed by the constrained gradient ascent algorithm in Section 3. The proposed supervised and semi-supervised approaches are introduced in Section 4 and 5 respectively. Experiments comparing different techniques are shown in Section 6, concluding remarks appear in Section 7.

2. Objective Function

Consider X, an $M \times N$ normalized term-document matrix, where M is the number of features and N is the number of documents. If we take the features W as a random variable over a finite set $\mathcal{W} = \{w_1, ..., w_M\}$, each document $\boldsymbol{x} \in X$ can be viewed as a probability distribution of the random variable W under a certain class C: $P(W|C)$. X is then interpreted as a probability matrix with the entry at position (i,j) being the probability $p_j(w_i|c)$. We reduce the dimensionality by projecting the features of each document onto a new feature space $\mathbb{R}^{\mathcal{W}'}$, via a transformation matrix $P = P(W'|W)$, where the (i,j)th entry is the probability of a new feature w'_j given the original feature w_i: $P(W' = w'_j|W = w_i)$.

After the projection we want the new features W' to retain as much information about the class labels C as possible. The mutual information [3] provides a quantitative measure of the mutual dependence between two random variables, which makes it a good choice for our objective function. Hence we seek a P which maximizes the mutual information between C and W': $I(C;W')$.

$$\begin{aligned} P &= \arg\max_{\tilde{P}} I(C;W') \\ &= \arg\max_{\tilde{P}} I(C;\hat{P}^TW) \end{aligned} \tag{1}$$

As shown above, the random variables C, W, W' form the following Markov Chain:

$$C \rightarrow W \rightarrow W' \tag{2}$$

Their relationships can be further illustrated using the information channel shown in Figure 1. According to the chain rule [3] we have $I(C;W) \geq I(C;W')$, $I(C;W)$ is the upper bound and can be easily estimated from labeled training data, where $P(C)$ and $P(W|C)$ are fixed and known. We will show in the following lemma that $I(C;W')$ is a convex function of

*The work was done when the third author was affiliated with Johns Hopkins University

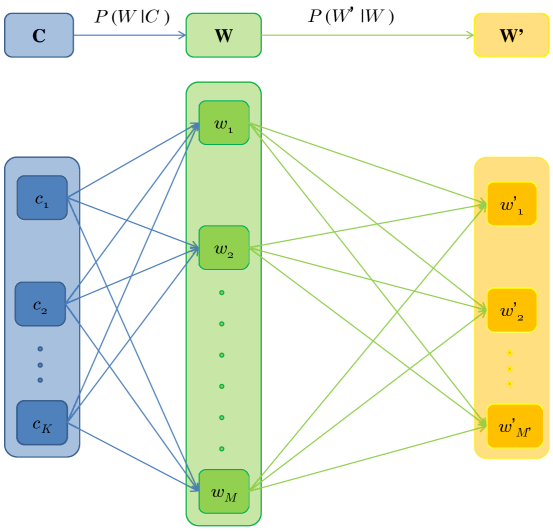


Figure 1: *Illustration of the Markov chain relations among the class label C, the high-dimensional feature W, the low dimensional feature W'*

$P(W'|W)$; this makes its maximization hard, as the maximum is achieved at one of the extremal points of the probability simplex. However, we are able to compute a local maximum based on the algorithm of [4].

Lemma 1 *$I(C;W')$ is convex with respect to $P(W'|W)$, assuming $C \to W \to W'$, $P(C)$ and $P(W|C)$ are fixed.*

Proof We first note that $I(C;W')$ is convex w.r.t $P(w'|c)$ with $P(c)$ fixed [3]. Also, since $C \to W \to W'$ forms a Markov chain, $P(w'|c,w) = P(w'|w)$.

$$\begin{aligned} P(w'|c) &= \textstyle\sum_w P(w',w|c) \\ &= \textstyle\sum_w P(w'|c,w)P(w|c) \\ &= \textstyle\sum_w P(w'|w)P(w|c) \end{aligned} \tag{3}$$

Since $P(w|c)$ is fixed, $I(C,W') = I(P(c),P(w'|c))$ can also be written as $I(C,W') = I(P(c),P(w'|w))$. Also:

$$\begin{aligned} P_\lambda(w'|c) &= \lambda P_1(w'|c) + (1-\lambda)P_2(w'|c) \\ &= \lambda \textstyle\sum_w P_1(w'|w)P(w|c) \\ &\quad + (1-\lambda)\textstyle\sum_w P_2(w'|w)P(w|c) \\ &= \textstyle\sum_w [\lambda P_1(w'|w) \\ &\quad + (1-\lambda)P_2(w'|w)]P(w|c) \\ &= \textstyle\sum_w P_\lambda(w'|w)P(w|c) \end{aligned} \tag{4}$$

We now have that $I(C;W')$ is convex w.r.t. $P(w'|w)$ because of the linear relationship between $P(w'|c)$ and $P(w'|w)$.

3. Constrained Gradient Ascent Method

Since the transformation matrix $P = P(W'|W)$ is a probability matrix, we are going to maximize the objective function $I(C;W')$ under the following constraints on the probability vector $\boldsymbol{p}_w = (p(w'_1|w),\cdots,p(w'_{M'}|w))$:

$$\textstyle\sum_{i=1}^{M'} p(w'_i|w) = 1 \tag{5}$$

$$p(w'_i|w) \geq 0 \tag{6}$$

We can simplify the above constraints by introducing an intermediate variable $\boldsymbol{l}_w$:

$$\boldsymbol{l}_w = \big(l(w'_1|w),\cdots,l(w'_{M'}|w)\big) \tag{7}$$

$$l(w'_i|w)^2 = p(w'_i|w) \tag{8}$$

The new constraint becomes:

$$\boldsymbol{l}_w \cdot \boldsymbol{l}_w^T = 1, \ \ \forall\, w \in W. \tag{9}$$

From Lemma 1 we can see that $I(C;W')$ is not a convex function w.r.t $\boldsymbol{l}_w$, instead it has multiple local optima. For a maximizing $\boldsymbol{p}_w$, there will be at most $2^{M'}$ different $\boldsymbol{l}$ corresponding to it, i.e. there are at most $2^{M'}$ different local optima for $\boldsymbol{l}$. However, since $\boldsymbol{l}_w^2 = \boldsymbol{p}_w$ these different local optima are all equivalent, and we can still employ the gradient approach to maximize $I(C;W')$ w.r.t. $\boldsymbol{l}_w$.

Since each row of P is a probability vector $\boldsymbol{p}_w$ of the new features W' for a given original feature $W = w$, the optimization process can be simplified to $|W| = M$ sub-steps. That is, in the i-th sub-step, we maximize $I(C;W')$ w.r.t. $\boldsymbol{p}_i$, which is the i-th row of the transformation matrix P. After all the M sub-steps are finished, we update the transformation matrix P, iterating this process until convergence.

4. Supervised CMMI

The optimization problem in each sub-step can be solved through the constrained gradient algorithm proposed by [4], and we use the following algorithm to find the maximizing $\boldsymbol{l}_w$. Accordingly we have the matrix L of the same size with the matrix P: $L^2 = P$.

Algorithm 1 Supervised CMMI

1. Initialize each row $\boldsymbol{l}_i$ of L with a nonzero random vector and normalize $\boldsymbol{l}_i$ s.t. $\boldsymbol{l}_i\boldsymbol{l}_i^T = 1$.
2. **For** $i = 1,2,\cdots,M$
   - Compute the direction vector: $\boldsymbol{h} = \frac{\partial I(C;W')}{\partial \boldsymbol{l}_i}$
   - Compute the moving step: $\alpha = -\left(\boldsymbol{h}^T \frac{\partial^2 I(C;W')}{\partial \boldsymbol{l}_i^2}\boldsymbol{h}\right)^{-1}\boldsymbol{h}^T\boldsymbol{h}$
   - Update $\boldsymbol{l}_i$ using $\boldsymbol{l}_i = \boldsymbol{l}_i + \eta\boldsymbol{h}|\alpha|$
   - "Normalize" the new $\boldsymbol{l}_i$ s.t. $\boldsymbol{l}_i\boldsymbol{l}_i^T = 1$

   End
3. Repeat step 2 until convergence.

where η is the step size. Since $I(C;W') = H(C) - H(C|W')$ and $H(C)$ is fixed, the derivatives of $I(C;W')$ w.r.t. $P(w'|w)$ can be written as follows:

$$\begin{aligned} &\frac{\partial I(C;W')}{\partial P(w'|w)} = -\frac{\partial H(C|W')}{\partial P(w'|w)} \\ &= \frac{\partial}{\partial P(w'|w)} \sum_{c,\hat{w}'} P(c,\hat{w}') \log P(c|\hat{w}') \\ &= \frac{\partial}{\partial P(w'|w)} \sum_{c,\hat{w}'} \left(\sum_{\hat{w}} P(\hat{w}'|\hat{w})P(\hat{w}|c)P(c)\right) \times \\ &\quad \log \frac{\sum_{\hat{w}} P(\hat{w}'|\hat{w})P(\hat{w}|c)P(c)}{\sum_{\hat{w}} P(\hat{w}'|\hat{w})P(\hat{w})} \\ &= \sum_c P(c,w)\log P(c|w') \end{aligned} \tag{10}$$

We can compute the derivative of $I(C;W')$ w.r.t. $l(w'|w)$ as follows:

$$\begin{aligned}\frac{\partial I(C;W')}{\partial l(w'|w)} &= \frac{\partial I(C;W')}{\partial P(w'|w)}\frac{\partial P(w'|w)}{\partial l(w'|w)} \\ &= 2l(w'|w)\sum_c P(c,w)\log P(c|w')\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial^2 I(C;W')}{\partial l(w'|w)^2} &= \frac{\partial}{\partial l(w'|w)}\, 2l(w'|w)\sum_c P(c,w)\log P(c|w') \\ &= \frac{\partial}{\partial l(w'|w)}\, 2l(w'|w)\times \\ &\qquad \sum_c P(c,w)\log\frac{\sum_{\hat{w}} P(c)P(\hat{w}|c)l^2(w'|\hat{w})}{\sum_{\hat{w}} P(\hat{w})l^2(w'|\hat{w})} \\ &= 2\sum_c P(c,w)\log P(c|w') + \\ &\quad 4\left[\sum_c P(c,w)\frac{P(c,w,w')}{P(c,w')} - \sum_c P(c,w)\frac{P(w,w')}{P(w')}\right]\end{aligned} \tag{12}$$

5. Semi-supervised CMMI

Incorporating additional information from unlabeled data can often improve performance, so we also explore a semi-supervised extension of CMMI through the following objective function:

$$(1-\gamma)I(C_{tr};W') + \gamma I(\widetilde{C}_{ts};W') \tag{13}$$

where C_{tr} is the true class label of the training data X_{tr}, and $\widetilde{C}_{ts}$ is the 'estimated' class labels of the unlabeled data[1] X_{ts} based on the results of supervised CMMI. Here, $\gamma \in [0,1]$ is the weight given to evidence from the unlabeled data.

In the following algorithm, W_{tr}, W_{ts} denote the data partitions, all corresponding to W in the high-dimensional space; similarly W'_{tr}, W'_{ts} correspond to W' in the low-dimensional space. X'_{tr} and X'_{ts} are the term document matrices of transformed training and test data respectively. The semi-supervised CMMI iteratively updates the transformation matrix $P^i = P^i(W'|W)$ as follows:

Algorithm 2 Semi-supervised CMMI

1. Use supervised CMMI to get $P^0 = P^0(W'_{tr}|W_{tr})$ and compute $X'_{tr} = {P^0}^T X_{tr}$, $X'_{ts} = {P^0}^T X_{ts}$
2. Train SVM on X'_{tr} and then apply it on X'_{ts} to get $\widetilde{C}^0_{ts}$.
3. **While** $\widetilde{C}^i_{ts} \neq \widetilde{C}^{i-1}_{ts}$
   - Find P^{i+1} that maximizes: $(1-\gamma)I(C_{tr};W'_{tr}) + \gamma I(\widetilde{C}^i_{ts};W'_{ts})$
   - Update $X'_{tr} = {P^{i+1}}^T X_{tr}$, $X'_{ts} = {P^{i+1}}^T X_{ts}$
   - Train SVM on X'_{tr} and then apply it on X'_{ts} to get $\widetilde{C}^{i+1}_{ts}$

   End
4. Output the converged $P = P(W'|W)$

[1] For our purposes, we only used unlabeled data from the test partition, though that is not necessary.

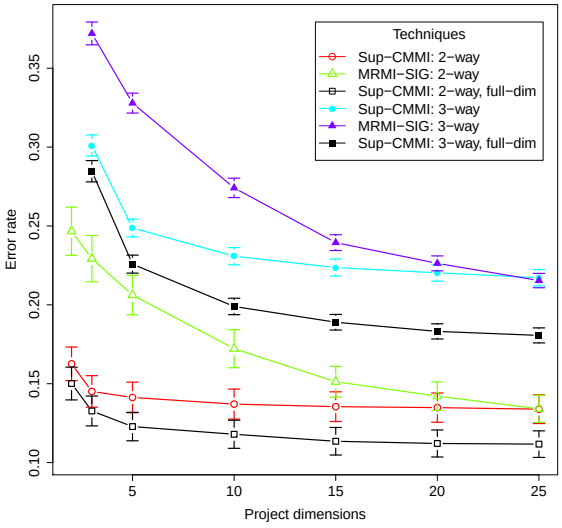


Figure 2: *Comparison of supervised CMMI and MRMI-SIG for 2-way and 3-way experiments*

6. Experimental Results

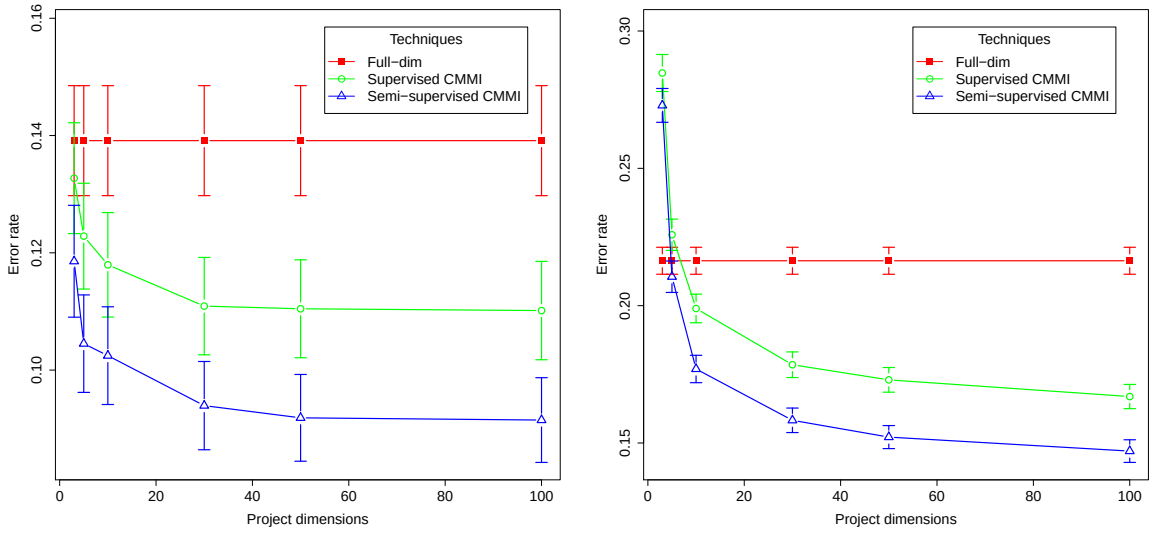
Our LID experiments are performed using the CallFriend corpus[2], which contains conversations in 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese. Each document is a 30-second segment of conversational telephone speech converted to a feature vector of dimension 12664, courtesy of [5]. For each language we have 800 documents randomly partitioned into training, development (for tuning γ in semi-supervised CMMI), and test (400, 200, and 200 documents respectively).

For each condition, we compute the average error rate μ and standard error s_e, with plots indicating $\mu \pm s_e$. Support vector machines (SVMs) with RBF kernel [6] were used for 2-way and 3-way classification, i.e. there are 2 classes and 3 classes respectively. All classifiers operate on the reduced-dimensionality features (W'), resultant from the application of CMMI or MRMI-SIG.

6.1. Supervised CMMI and MRMI-SIG

We compare CMMI and MRMI-SIG [2] (considered state-of-the-art), on 2- and 3-way LID tasks. Unfortunately, performing MRMI-SIG on the full dimensionality of the data (12664 dimensions) is computationally prohibitive, so we are forced to make two sets of comparisons to adequately characterize their relative performance. First, we compare CMMI and MRMI-SIG on equal footing by performing a standard dimensionality reduction technique (principal components analysis; PCA) on the data to 30 dimensions, then comparing their performances on reducing the dimensionality further. This comparison is biased in favor of MRMI-SIG, since we remove one of the primary advantages of CMMI – using the natural (high) dimensionality of the data. Second, we remove this artificial restraint and allow CMMI to work on the full dimensionality, rather than the 30 dimensional representation resultant from PCA.

[2] `http://www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone-conversations`



(a) (b)

Figure 3: *Error rates of supervised and semi-supervised CMMI: (a) 2-way classification; (b) 3-way classification*

For all experiments, the average classification error rates μ across all possible language mixtures are shown in Figure 2. CMMI generally has a lower average error rate than MRMI-SIG, especially in low projection dimensions; for $d \leq 15$ the improvement is statistically significant ($p \leq 1.2 \times 10^{-6}$). As expected, when CMMI operates on the full dimensionality of the data, further improvements in performance are evident.

6.2. Supervised and semi-supervised CMMI

We compare the performance on 2- and 3-way classification tasks on the full dimensionality to those in reduced dimensionality using either supervised CMMI or semi-supervised CMMI. For semi-supervised CMMI, we optimize γ on the development set for each language mixture, selecting from $\gamma \in \{0, 0.25, 0.5, 0.75, 0.9, 1\}$.

We project the data into $\mathbb{R}^d$ for $d \in \{3, 5, 10, 30, 50, 100\}$. Figure 3 indicates that for 2-way classification, both the supervised and semi-supervised CMMI give lower average error rates μ than the full-dimensionality case. This difference is statistically significant when $d \geq 5$, ($p \leq 4.2 \times 10^{-7}$); for 3-way experiments, when $d \geq 10$ they both have a lower μ ($p \leq 2.2 \times 10^{-16}$). In all cases semi-supervised CMMI is also statistically significant compared against supervised CMMI ($p \leq 2.7 \times 10^{-4}$).

Two key observations from these results are in order: (i) Dimensionality reduction is crucial for tackling the curse of dimensionality, even when the classifier is a SVM. As our results demonstrate, our particular dimensionality reduction technique results in significant gains over using the full dimensionality on this dataset. (ii) Semi-supervised learning, that utilizes automatically generated class labels, results in further gains over using supervised data only.

7. Conclusion

In this paper we propose CMMI, an information-theoretic dimensionality reduction technique that maximizes the mutual information between the class labels C and the projected features in lower dimensional space W'. CMMI views the feature vector of each document $\boldsymbol{x}$ as a probability vector and finds a matrix $P = P(W'|W)$ to perform the projection: $\boldsymbol{x}' = P^T\boldsymbol{x}$. We demonstrate superior classification performance on a standard LID task as compared to state-of-the-art techniques. Additionally, we also demonstrate scalability to higher dimensional data than the state-of-the-art can operate on.

8. Acknowledgements

The first and third authors would like to acknowledge support by the National Science Foundation grant No CCF-0728931.

9. References

[1] K. Bollacker and J. Ghosh, "Mutual information feature extractors for neural classifiers," in *International Conference on Neural Networks*, 1996, pp. 1433–1441.

[2] K. T. K.E. Hild II, D. Erdogmus and J. Principe, "Sequential feature extraction using information theoretic learning," in *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, 2006, pp. 1385–1393.

[3] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., Hoboken, NJ. 2006.

[4] A. A. Hasan and M. A. Hasan, "Constrained gradient descent and line search for solving optimization problem with elliptic constraints," in *ICASSP*, vol. 2, 2003, pp. 793–796.

[5] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Proc. of Interspeech*, Brighton, UK, 2009.

[6] K. H. A. Karatzoglou, A. Smola and A. Zeileis, "kernlab – an S4 package for kernel methods in R," in *Journal of Statistical Software*, vol. 11, no. 9, 2004.