

Low-rank Audio Signal Classification Under Soft Margin and Trace Norm Constraints

Ziqiang Shi, Tieran Zheng, Jiqing Han, Shiwen Deng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

{zqshi, zhengtieran, jqhan}@hit.edu.cn, dengswen@gmail.com

Abstract

We propose an algorithm to do speech/non-speech classification based on low-rank matrix representative audio data. Conventionally, the low-rank matrix data can be represented by a vector in high dimensional space. Some learning algorithms are then applied in such a vector space for matrix data classification. Particularly, maximum margin classifiers, such as support vector machine (SVM) etc. have received a lot of attentions due to their effectiveness. In this paper, we classify the data directly in the matrix space. Our methodology is built on recent studies about matrix classification with the trace norm constrained weight matrix and SVM's large-margin linear discrimination principle. The resulting low-rank SVM is then designed to maximize the margin between classes whilst minimizing the complexity of the classifier in both original and low-rank space. We compared our proposed algorithm with SVM and other state-of-the-art matrix classification methods. Experimental studies on real life audio signal classification show the effectiveness of our algorithm.

Index Terms: speech/non-speech, matrix classification, maximum margin, trace norm regularization, low-rank feature

1. Introduction

Audio classification plays an important role in multimedia information retrieval, human-computer interfaces, entertainment, and surveillance. In speech processing systems developed for real world applications, a discrimination of speech from other audio classes is a crucial step. Speech/non-speech audio classification can also be used for coding or telecommunication applications where non-speech audio are not of interest, and hence bandwidth can be saved by not transmitting them or by assigning them a low resolution code. Finally, as the amount of available real-world audio data increases rapidly, such as Freesound [1], manual annotation of audio sounds has become more difficult and impractical, so alternative automated recognition to which class an audio sound belongs are much needed.

Unlike traditional vector form audio features for audio applications [2, 3], low-rank matrix features which are more robust to large corruptions than traditional features have been proposed for robust audio classification recently [4]. For such matrix descriptive features, various machine learning methods can be used to provide a final classification, such as rule-based approaches, Gaussian mixture models, support vector machines, Bayesian networks, etc. [3, 5]. These vector feature based machine learning methods can not fully make use of the spatial/correlation information embedded in the 2-D structure of the matrix data and also causes the "curse of dimensionality" problem. Different from these vector data based methods, re-

cently Tomioka and Aihara proposed the trace norm regularization framework based matrix classification in [6].

The goal of the framework is to infer the weight matrix and bias under low trace norm constraints and low deviation of the empirical statistics from their predictions. The trace norm was used to measure the complexity of the weight matrix of the linear classifier for matrix classifications. This kind of inference task belongs to the more general problem of learning low-rank matrix through convex optimization, and emerged in various applications recently, such as matrix completion [7], multi-task learning [8], and robust principle component analysis (robust PCA) [9, 10]. For the matrix rank minimization is NP-hard in general due to the combinatorial nature of the rank function, a commonly-used convex relaxation of the rank function is the trace norm (nuclear norm) [11], defined as the sum of the singular values of the matrix. This framework has been modified and adapted by Shi et al. [4] for audio data classification. Shi et al. proposed the learning problem as a minimization of the combination of the trace norm and the ℓ_2 -norm.

Generally these above regression methods have a low performance compared to SVM in theory and applications for vector based applications. SVM is based on the Structural Risk Minimization (SRM) principle for which error-bound analysis has been theoretically motivated. The method is defined over a vector space where the problem is to find a decision surface that maximizes the margin between the data points in a training set. In this paper, we generalize SVM to the case of matrix data classification with trace norm constraint. This novel generalization find a classifier in matrix space which leads to structured classification, while traditional SVM find a classifier in a vector space. It uses SVM's SRM principle to achieve almost best decision surface, at the same time uses both trace norm and Frobenius norm to control the complexity of the model from different space, thus the classifier obtained by solving this formulation is referred as low-rank SVM.

The rest of this paper is organized as follows: Section 2 provides a brief description of trace norm regularized matrix classification and maximum margin principle based classifiers. In Section 3, we give some descriptions on matrix classification under soft margin and trace norm constraints. The experimental results on speech/non-speech audio classification are presented in Section 4. Finally, we provide some concluding remarks and suggestions for future work in Section 5.

2. Preliminary

In this section, we provide a brief overview of spectral regularized matrix classification. For a detailed treatment please see [6, 12, 4].

2.1. Matrix Classification with Trace Norm Regularization

We consider linear matrix classification model:

$$y_i = \text{Tr}(W^T X_i) + b, \quad (1)$$

where $W \in \mathbb{R}^{m \times n}$ is the unknown *weight matrix*, $b \in \mathbb{R}$ is the *bias*, and $(X_i, y_i) \in \mathbb{R}^{m \times n} \times \mathbb{R} (i = 1, \dots, s)$ are the training samples. The weight and bias can be simply inferred by just ignoring the fact that the input is a matrix and concatenate all columns into a long vector

Tomioka and Aihara [6] introduce the trace norm to this task, and formulate the inferring as a convex optimization problem:

$$\begin{aligned} \min_{W, b} \quad & \sum_{i=1}^s \log(1 + \exp(-z_i)) + \lambda \|W\|_* \\ \text{subject to} \quad & y_i(\text{Tr}(W^T X_i) + b) = z_i \\ & \forall i \in \{1, \dots, s\}, \end{aligned} \quad (2)$$

where $\|\cdot\|_*$ denotes the trace norm of a matrix which is defined as the sum of the singular values.

Shi et al. [4] proposed an alternative approach for low-rank audio data classification:

$$\min_{W, b} \sum_{i=1}^s (y_i - \text{Tr}(W^T X_i) - b)^2 + \lambda \|W\|_*. \quad (3)$$

All these methods have efficient performance in each application. In supervised learning settings with many input features, previous investigations have demonstrate that SVM can be superior to other learning algorithms. Thus the principles used in SVM may be adopted in our matrix classification task.

2.2. SVM

The central idea of classical SVM is to separate data by finding a vector $W \in \mathbb{R}^{m \times n}$ and an offset $b \in \mathbb{R}$ of a hyperplane which provides us with the largest possible margin between classes and minimizes the number of misclassified patterns. If do not distinguish between the variables, the optimization problem of SVM for matrix classification can be stated as follows:

$$\begin{aligned} \min_{W, b} \quad & \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i(\text{Tr}(W^T X_i) + b) \geq 1 - \xi_i \quad \forall i \in \{1, \dots, s\} \\ \text{and} \quad & \xi_i \geq 0 \quad \forall i \in \{1, \dots, s\}, \end{aligned} \quad (4)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This formulation does not treat the data as 2-D matrix form and it essential identical to the vector based SVM. We can introduce spatial/correlation information into Equation (5) to control the complexity of the model and to make the formulation into a 2-D SVM.

3. Maximum Margin Matrix Classification With Spectral Constraints

Combining the spectral regularized matrix classification (Equation (2) and Equation (3)) and principles of SVM (Equation (5)), we obtain the trace norm regularized maximum margin matrix classification:

$$\begin{aligned} \min_{W, b} \quad & \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^n \xi_i + \lambda \|W\|_* \\ \text{subject to} \quad & y_i(\text{Tr}(W^T X_i) + b) \geq 1 - \xi_i \\ \text{and} \quad & \xi_i \geq 0, \lambda > 0 \\ & \forall i \in \{1, \dots, s\}. \end{aligned} \quad (5)$$

By introducing variable Z , we can rewrite the trace norm with Equation (5) as follows:

$$\begin{aligned} \min_{W, Z, b} \quad & \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^n \xi_i + \lambda \|Z\|_* \\ \text{subject to} \quad & y_i(\text{Tr}(W^T X_i) + b) \geq 1 - \xi_i \\ \text{and} \quad & \xi_i \geq 0, \lambda > 0 \\ & \forall i \in \{1, \dots, s\} \\ \text{and} \quad & W = Z. \end{aligned} \quad (6)$$

Furthermore, the equation constraint $W = Z$ can be relaxed to $\|W - Z\|_F < \varepsilon$ for robustness, and the problem can be equivalently transformed into

$$\begin{aligned} \min_{W, Z, b} \quad & \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^n \xi_i + \lambda \|Z\|_* + \frac{v}{2} \|W - Z\|_F^2 \\ \text{subject to} \quad & y_i(\text{Tr}(W^T X_i) + b) \geq 1 - \xi_i \\ \text{and} \quad & \xi_i \geq 0, \lambda > 0, v > 0 \\ & \forall i \in \{1, \dots, s\}. \end{aligned} \quad (7)$$

In this paper, we focus on this formulation of the problem. The objective function is not convex with respect to all parameters jointly, but is in each of its parameters. Hence, we proceed by alternating the search for a solution of (7). For Z fixed and let $\tilde{W} = W - \frac{v}{v+1}Z$, $\tilde{\xi}_i = \xi_i + y_i \frac{v}{v+1} \text{Tr}(Z^T X_i)$, then the problem (7) can be formulated as follows:

$$\begin{aligned} \min_{\tilde{W}, b} \quad & \frac{v+1}{2} \|\tilde{W}\|_F^2 + C \sum_{i=1}^n \tilde{\xi}_i \\ \text{subject to} \quad & y_i(\text{Tr}(\tilde{W}^T X_i) + b) \geq 1 - \tilde{\xi}_i \\ \text{and} \quad & \tilde{\xi}_i \geq -y_i \frac{v}{v+1} \text{Tr}(Z^T X_i) \\ & \forall i \in \{1, \dots, s\}, \end{aligned} \quad (8)$$

which is a ℓ_1 soft margin SVM with the only difference being that 0 is replaced by $-y_i \frac{v}{v+1} \text{Tr}(Z^T X_i)$ in the inequality constraint. The primal and dual objective functions of such a problem are convex, and their solution can be obtained by any of the available SVM algorithms. Here we opt for LIBSVM [13].

On the other hand, for W fixed, the problem is reduced to

$$\min_{Z, b} \lambda \|Z\|_* + \frac{v}{2} \|W - Z\|_F^2. \quad (9)$$

According to the Theorem 3.1. of [14], Equation (9) can be solved by first computing the singular value decomposition (SVD) of W and then applying some soft-thresholding on the singular values. For details please refer Algorithm (1), in which we give a whole outline of solving Equation (7). The $\mathcal{S}_\varepsilon[\cdot]$ in Algorithm (1) is the soft-thresholding operator introduced in [10]:

$$\mathcal{S}_\varepsilon[x] \doteq \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon, \\ x + \varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where $x \in \mathbb{R}$ and $\varepsilon > 0$.

We initialize $W = Z$ as $\min(m, n)^{-1}I$, where $I \in \mathbb{R}^{m \times n}$ with $\min(m, n) \times \min(m, n)$ identity matrix in the right up corner. We proceed by alternating the search for a solution of (7) for Z , given a fixed W, b and for W, b given previous estimation of Z . The alternating optimization scheme is stopped when the Frobenius norm of the changes on Z becomes negligible.

4. Experimental Validation

In this section, we conduct experiments on real audio data to demonstrate the characteristics of the proposed methods for

Algorithm 1 Learning of low-rank SVM classifier

Input $(X_i, y_i), i = 1, \dots, s$.**Initialization** $W_0 = Z_0 = \min(m, n)^{-1} I$, where $I \in \mathbb{R}^{m \times n}$ with $\min(m, n) \times \min(m, n)$ identity matrix in the right up corner, $k = 1$.1: **while** not converged **do**2: Obtain W_k, b_k by solving Equation (8) using LIBSVM with $Z = Z_{k-1}$.3: $(U, S, V) = \text{svd}(W_k)$.4: $Z_k = U S_{\Delta}^v [S] V^T$.5: $k \leftarrow k + 1$.6: **end while****Output:** $Z \leftarrow Z_k, b \leftarrow b_k$.

tensor classification problem. Three algorithms are compared: Tomioka and Aihara's method [6] (TA); Shi et al.'s method (Shi); the trace norm regularized maximum margin matrix classification (Low-rank SVM) proposed in this paper. All algorithms are run in Matlab on a personal computer with an Intel 3.40 GHz dual-core central processing unit (CPU) and 2GB memory. The regularization constant λ is set $1/\sqrt{650}$ which is a classical normalization factor according to [15]. v is set 1, which are determined by empirical evidence and the automatic determination should be further studied, but not in this work.

4.1. Speech/non-speech audio classification

Experiments are conducted on a collected database, which is the one used in [16]. The audio database was assembled from public corpora and the Freesound project [1]. Speech samples were taken from GTZAN music/speech dataset¹, Scheirer&Slaney music/speech corpus² [17], and emoDB³ [18]. 600 speech samples were selected from these three datasets. For non-speech class, we downloaded 2512 audio samples from the Freesound project [1] by using the most popular tags, including "ambience", "bass", "digital", "drum", "electronic", "industrial", "machine", "music", "nature", "piano", "rain", "street", "train", and "water". For these downloaded audio recordings, we first manually filtered them to discard audio that contains much speech, or low-quality audio. In consequence, we select 1,800 non-speech samples. All the audio recordings were converted to monaural wave format at a sampling frequency of 8kHz and quantized 16bits. Furthermore, the audio signals have been normalized, so that they have zero mean amplitude with unit variance in order to remove any factors related to the recording conditions.

Audio streams were windowed into a sequence of short-term frames (20 ms long) with no overlap. Then adjacent 50 frame signals are made together as rows, then the audio segments become matrices. Further these matrices are mapped into a low-rank matrix space. In this work we transform the recovered low-rank matrices into MFCCs (mel-frequency cepstral coefficients) [19] matrices, that means all rows in the low-rank matrices are transformed into MFCCs independently. Fig. 1 shows the low-rank matrix and the spectrograms of a typical speech sound effect audio segment. For details of low-rank audio feature extraction please refer [4]. The goal here is to classify the matrices according to their labels. Speech/non-speech classification learning task is used to evaluate the performance

of the online learning method.

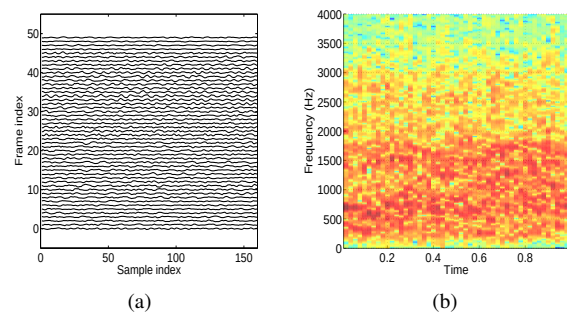


Figure 1: Low-rank component of the audio segment and the corresponding spectrogram of the low-rank component. (a) The low-rank component recovered from the matrix form of the speech segment; (b) Spectrograms of the low-rank approximation of the speech segment.

We first show that Algorithm 1 is convergent experimentally using the two learning tasks. In theory, the two steps for solving W_k and Z_k in each iteration are solutions of two convex sub-problems, thus the algorithm is indeed convergent. Fig. 2 shows the values of the target functions as functions of time for the two tasks respectively. It can be seen that generally the low-rank SVM is convergent after 15-25 iterations for our tasks.

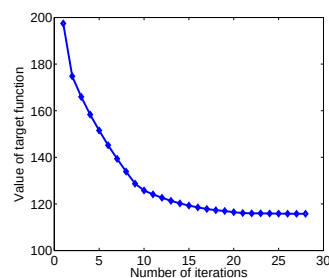


Figure 2: Convergence of the target function values in learning low-rank SVM for speech/non-speech classification.

Fig. 3 compares the three algorithms. Classifiers are trained on 5/6 of the data and then tested on the remaining 1/6. It shows the classification rates on testing data as functions of time. We can observe that the low-rank SVM is by far the most efficient method in speech/non-speech discrimination. Table 1 reports the classification rates after all the algorithms converged. Low-rank SVM obtains the best classification accuracies on both tasks. We also compare the spectral regularized methods with the state-of-the-art SVM classifier with long vector feature (650 dimension) obtained by vectorizing the low-rank matrix feature. It can be seen that may be due to the information lost in vectorizing the low-rank matrix feature, all spectral regularized matrix classification methods outperform SVM.

5. Conclusions

In this paper, low-rank features are introduced for speech/non-speech classification and spectral norm regularized matrix classification are introduced in the SVM theoretical framework.

¹http://marsyas.info/download/data_sets.

²<http://labrosa.ee.columbia.edu/sounds/musp/scheislan.html>.

³<http://pascal.kgw.tu-berlin.de/emodb/>.

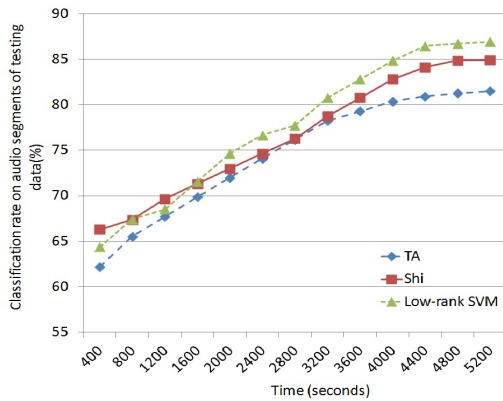


Figure 3: Comparisons of the methods in speech/non-speech segments classification.

Table 1: Speech/non-speech classification comparison between our approach and previous methods.

	TA	Shi	Low-rank SVM	SVM
Precision	81.04%	86.27%	88.01%	77.36%

Hence, it finds a classifier with low complexity in the low-rank space, and the same classifier also maximize the margin between classes. The proposed method guarantees an automatic and reliable determination of the optimal surface in the low-rank classifier space. Though the low-rank SVM was designed as a classifier for audio classification applications, there is no reason to restrict it for this particular application. Experimental results on low-rank audio signal classification illustrate the efficiency of the method. The algorithm performs well in terms of speech/non-speech classification. However, the optimization problems of the method is not convex; the classifier training procedure may converges to a local minimum. In future work, we plan to proposes a convex framework for jointly learning optimal W and Z .

6. Acknowledgement

This work was partly supported by the National Natural Science Foundation of China under grant No. 91120303 and No. 61071181.

7. References

- [1] <http://www.freesound.org>, “The freesound project.”
- [2] P. Atrey, N. Maddage, and M. Kankanhalli, “Audio based event detection for multimedia surveillance,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.
- [3] K. Umaphathy, S. Krishnan, and R. Rao, “Audio signal feature extraction and classification using local discriminant bases,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1236–1246, 2007.
- [4] Z. Shi, J. Han, and T. Zheng, “A novel framework based on trace norm minimization for audio event detection,” in *Proceedings of the 18th International Conference On Neural Information Processing*. Springer, 2011, pp. 646–654.
- [5] X. Zhuang, X. Zhou, T. Huang, and M. Hasegawa-Johnson, “Feature analysis and selection for acoustic event detection,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 17–20.
- [6] R. Tomioka and K. Aihara, “Classifying matrices with a spectral regularization,” in *Proceedings of the 24th International Conference on Machine Learning*. ACM, 2007, pp. 895–902.
- [7] N. Srebro, J. Rennie, and T. Jaakkola, “Maximum-margin matrix factorization,” *Advances in neural information processing systems*, vol. 17, pp. 1329–1336, 2005.
- [8] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex multi-task feature learning,” *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.
- [9] J. Wright, A. Ganesh, S. Rao, and Y. Ma, “Robust principal component analysis?” in *Proceedings of the Conference on 2009 Neural Information Processing Systems*. Citeseer.
- [10] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” *Arxiv preprint arXiv:1009.5055*, 2010.
- [11] M. Fazel, H. Hindi, and S. Boyd, “A rank minimization heuristic with application to minimum order system approximation,” in *Proceedings of the 2001 American Control Conference.*, vol. 6. IEEE, 2001, pp. 4734–4739.
- [12] V. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.
- [13] C. Chang and C. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [14] S. Ji and J. Ye, “An accelerated gradient method for trace norm minimization,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 457–464.
- [15] P. Bickel, Y. Ritov, and A. Tsybakov, “Simultaneous analysis of lasso and dantzig selector,” *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.
- [16] Z. Shi, J. Han, and T. Zheng, “Real-world speech/non-speech audio classification based on sparse representation features and gpcs,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [17] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1331–1334.
- [18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [19] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.