International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 20, No. 3 (2006) 431–440
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

LEARNING SPARSE MIXTURE MODELS FOR DISCRIMINATIVE CLASSIFICATION

WEIXIANG LIU* and NANNING ZHENG†

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
Xi'an,Shaanxi Province 710049, P. R. China
*victorwxliu@yahoo.com.cn
†nnzheng@mail.xjtu.edu.cn

SONGFENG ZHENG

Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095, USA
sfzheng@stat.ucla.edu

Recently Saul and Lee proposed a mixture model for discriminative classification of non-negative data via non-negative matrix factorization for feature extraction. In order to improve the generalization, this paper considers a sparse version of the model. The basic idea is to minimize the sum of the weights of un-normalized mixture models for *posterior* distributions according to regularization method. Experiments on CBCL face database and USPS digit data set assess the validity of the proposed approach.

Keywords: Discriminative classification; mixture model; sparseness; nonnegative matrix factorization; regularization method.

1. Introduction

Mixture models have been widely investigated for classification. There exist two strategies for training a mixture model: generative (also called informative) learning and discriminative learning. The former is an indirect method, which firstly calculates the class conditional distribution and then computes the posterior distribution via Bayesian rule; the latter is a direct one, which models the boundaries or the posterior directly; see Refs. 6 and 17 for more detail.

This paper mainly focuses on the discriminative learning based on Saul and Lee's work,[19] where they investigated a learning algorithm using mixture models for the classification of non-negative data, but the results indicate the method does not have good generalization. Starting with analyzing the discriminative objective function, we show its relationship with maximum likelihood and maximum entropy principle, and that poor generalization will occur without *prior* knowledge. Regularization is a powerful tool to improve generalization, and has been successfully applied in many learning tasks, e.g. support vector machines,[22] AdaBoost.[15] Based on regularization theory, this paper proposes a sparse mixture model, whose basic idea is to impose sparseness constraints on the mixture weights of the *posterior* distributions. It is the only difference with Saul and Lee's method.

The rest of this paper is organized as follows. After a short review on learning and generalization in Sec. 2, Sec. 3 proposes the sparse mixture models for discriminative learning based on regularization method along with the discussion of related methods in Sec. 4; the experimental results on CBCL face database and USPS digit data set are given in Sec. 5, and we conclude the paper in Sec. 6.

2. Review on Learning and Generalization

The task of classification is to design a classifier based on N given *i.i.d.* samples $\{x_n, y_n\}, n = 1, 2, \ldots, N$, where $x_n \in R^d$ and $y_n \in \{1, 2, \ldots, I\}$ for I categories. Many algorithms have been proposed for this problem, from statistical methods to neural networks, e.g. see Refs. 3, 4, 16, and the recent support vector machines.[22]

Theoretically, we can use the joint pdf $p(x, y, \Theta)$ for data and parameter Θ (also contains some structure information) to describe the rule underlying the observation.[13] We may consider the following simple factorization of the joint pdf as

$$p(x, y, \Theta) = p(x|y, \Theta)p(y|\Theta)p(\Theta) \tag{1}$$

$$= p(y|x, \Theta)p(x|\Theta)p(\Theta) \tag{2}$$

where $p(\Theta)$ stands for some *prior* knowledge on the parameter Θ.

Now we define the likelihood function based on the N training samples and parameter as

$$L = \prod_{n=1}^{N} p(x_n, y_n, \Theta). \tag{3}$$

Furthermore, it can be expressed, according to Eqs. (1) and (2), as

$$L = \prod_{n=1}^{N} p(x_n|y_n, \Theta)p(y_n|\Theta)p(\Theta) \tag{4}$$

$$= \prod_{n=1}^{N} p(y_n|x_n, \Theta)p(x_n|\Theta)p(\Theta). \tag{5}$$

When we want to maximize the functions above with respect to Θ, we can maximize their corresponding log functions as

$$\begin{aligned} E &= \log L \\ &= \sum_{n=1}^{N} \log p(x_n|y_n, \Theta) + \sum_{n=1}^{N} \log p(y_n|\Theta) + \sum_{n=1}^{N} \log p(\Theta) \qquad (6) \\ &= \sum_{i=1}^{N} \log p(y_n|x_n, \Theta) + \sum_{n=1}^{N} \log p(x_n|\Theta) + \sum_{n=1}^{N} \log p(\Theta). \qquad (7) \end{aligned}$$

Now we can see that the common maximum likelihood estimation is to maximize (6) without consideration of $p(\Theta)$. Here we focus more on Eq. (7): the first term is the sum of *posterior* over all training samples, and the last two terms stand for some *prior* background information on x_i, y_i and the parameter Θ, which always play a role of regularization term according to the selection of *prior* knowledge, e.g. see Ref. 13. In fact, this can be considered as one kind of the maximum *a posteriori* estimation.

When the *prior* knowledge related to x_i, y_i, Θ is not taken into account, we get the following two common functions for maximization

$$E_{\text{gen}}(\Theta) = \sum_{n=1}^{N} \log p(x_n|y_n, \Theta) \tag{8}$$

and

$$E_{\text{dis}}(\Theta) = \sum_{n=1}^{N} \log p(y_n|x_n, \Theta). \tag{9}$$

Usually speaking, Eq. (8) is used for generative learning and Eq. (9) for discriminative learning.

On the other hand, according to maximum entropy principle, we can define a loss function for maximization as[8]

$$E_{\text{MaxEnt}}(\Theta) = -\sum_{n=1}^{N} p(y_n|x_n, \Theta) \log p(y_n|x_n, \Theta). \tag{10}$$

It is shown that maximizing Eq. (9) equals to maximizing Eq. (10),[8] based on the maximum entropy principle. However, many previous results show that maximum entropy principle is only an inference process with some common sense principles of uncertain reasoning; see Ref. 14, and the references therein. So we should add more constraints on Eq. (9) to get the optimal resolution when dealing with some specified tasks such as object recognition. Because of no *prior* knowledge from special domain, it leads to poor generalization based on only optimizing Eqs. (8) or (9). In practice, there are many methods to improve generalization,[18] and regularization method is one popular approach; see Ref. 2 for a good survey. According to the regularization method, we will discuss a spare mixture model for discriminative learning based on Saul and Lee's algorithm[19] in the next section.

3. Sparse Mixture Models for Discriminative Classification

When applying generative mixture models for learning, we first model the class density distribution. In this case, the class density distribution with parameter $\Theta = \{\Theta_1, \ldots, \Theta_M\}$ may take the following form as[19]

$$p(x|y = i, \Theta) = \sum_{j=1}^{M} W_{ij} \Phi(x|\Theta_j), \tag{11}$$

where M is the number of components for each class; and the mixture weights W_{ij} and basis functions $\Phi(x|\Theta_j)$ are all usually normalized, i.e. subjected to $\sum_j W_{ij} = 1$ for all i and $\int \Phi(x|\Theta_j)dx = 1$ for all j, respectively.

When mixture models are used for discriminative learning, the modeled *posterior* distributions can take the following form[19]

$$p(y = i|x, \Theta) = \frac{\sum_j W_{ij}\Phi(x|\Theta_j)}{\sum_{kl} W_{kl}\Phi(x|\Theta_l)}, \tag{12}$$

where the mixture weights W_{ij} and basis functions $\Phi(x|\Theta_j)$ are all nonnegative. The nonnegative constraints are derived from nonnegative data analysis [10] and we can make use of multiplicative updates.[9] For sparse nonnegative features, the basis function has the following form[19]

$$\Phi(x|\Theta_j) = e^{\Theta_j \cdot X}, \tag{13}$$

where Θ_j is a real vector and X denotes a nonnegative representation of the feature vector x.[19] The final objective function for discriminative training is to maximize the conditional log likelihood (9), which can be rewritten here as

$$L_c = \sum_n \log p(y_n|x_n, \Theta). \tag{14}$$

In fact we can impose some constraints, via some *prior* knowledge, on the unnormalized weights and basis functions. Here we consider one simple case: we hope that the weights are sparse, and this can be expressed via minimizing $\sum_{ij} W_{ij}$. We call this a *sparse* mixture model, and then we can get the objective function for discriminative training as below via regularization method

$$L_r = L_c - \lambda \sum_{ij} W_{ij}, \tag{15}$$

where λ is a regularization parameter with nonnegative value. The last term can be interpreted as imposing Laplace *prior* on W; and some more can be considered.[13,23]

Based on the learning algorithm of Saul and Lee's work,[19] it is easy to derive the multiplicative updating rules as following

$$W_{ij} \leftarrow W_{ij} \left\{ \left(\frac{\partial L_+}{\partial W_{ij}} \right) \Big/ \left(\frac{\partial L_-}{\partial W_{ij}} + \lambda \right) \right\} \tag{16}$$

$$e^{\Theta_{ju}} \leftarrow e^{\Theta_{ju}} \left\{ \left(\frac{\partial L_+}{\partial \Theta_{ju}} \right) \Big/ \left(\frac{\partial L_-}{\partial \Theta_{ju}} \right) \right\}^s \tag{17}$$

where

$$L_+ = \sum_n \log \sum_{ij} Z_{ni} W_{ij} e^{\Theta_j \cdot X_n}, \tag{18}$$

$$L_- = \sum_n \log \sum_{ij} W_{ij} e^{\Theta_j \cdot X_n}; \tag{19}$$

Z_{ni} is a binary matrix in which *ni*th element declares whether the *n*th training sample belongs to the *i*th class; and

$$s = \frac{1}{\max_n \sum_u X_{nu}} \tag{20}$$

is a measurement of the sparseness of the training features. See Appendix A for the simple and easily understood proof of convergence.

In order to get nonnegative sparse features, we adopt the nonnegative matrix factorization method which has been discussed in Refs. 9 and 10. However, how to select an optimal λ is still an open problem although there exist some methods.[2] We choose the values for λ heuristically in this paper. When $\lambda = 0$, the learning algorithm reduces to that in Ref. 19 without spareness constraints on weights.

4. Related Methods

Our method is based on regularization theory,[2] which has its Bayesian interpretation with Laplace *prior* on weights.[13,23] The basic idea has been used for supervised and unsupervised learning, e.g. see Refs. 5, 11, 12, 25 recently. In fact, according to the idea in Ref. 8, the proposed method here can also be used for speech recognition[20] and suchlike.

5. Experimental Results

5.1. *Binary classification*

Firstly, we tested our algorithm on the widely used CBCL face and nonface database[1] for binary classification. The data set contains a training set of 2,429 faces and 4,548 nonfaces, a test set of 472 faces and 23,573 nonfaces. The size of each gray image is $19 \times 19 = 361$ and all pixel values are between 0 and 1.

For computational efficiency, we used 2,429 faces and 2,500 nonfaces in the training set for learning, the first 5,000 samples of test set for test, which include 472 faces and 4,528 nonfaces. Nonnegative matrix factorization[10] was applied to discover sparse features with lower dimensionality, and we set $d = 80$ heuristically here. We also set five values for λ for comparison: 0, 0.1, 0.5, 1 and 5. We considered different models with different mixture components for each class: $M = 8, 16, 24, 32, 48$. All models were initialized with $W_{ij} = 0.5$ and $\Theta_{ju} = 0.5$ where the dimension of each Θ_j is equal to d, and trained by the same 100 iterations to converge for comparison.

As shown in Table 1, we can see that when the number of mixture components (M) per class increases, the accuracy of classification for face is improved. More importantly, the proposed approach with positive value for λ can improve the generalization in contrast to the original method, i.e. the case of $\lambda = 0$, especially for the class of face. For example, the highest accuracy rate for the class of face (F) is 96.8% when $M = 48$ and $\lambda = 5$; and that for the class of nonface (NF) is 74.1%

Table 1. Classification accuracy rates (%) on a test set of CBCL face database where F stands for the class of face and NF stands for the class of nonface.

	$\lambda = 0$		$\lambda = 0.1$		$\lambda = 0.5$		$\lambda = 1$		$\lambda = 5$	
M	F	NF	F	NF	F	NF	F	NF	F	NF
8	66.5	51.7	67.4	52.2	69.5	52.9	72.0	53.6	74.8	56.4
16	75.2	49.9	77.1	50.9	79.0	53.1	80.9	56.0	82.8	58.6
24	75.2	72.2	79.7	73.0	82.2	**74.1**	83.9	74.0	86.4	73.1
32	76.2	61.4	82.2	62.2	87.5	61.4	90.5	60.7	92.0	60.1
48	85.6	57.8	89.6	58.9	93.2	58.8	93.9	58.0	**96.8**	57.7

when $M = 24$ and $\lambda = 0.5$. The limitation of this method is how to select optimal λ and M for the best generalization, which is still an open problem.

5.2. *Multiclass classification*

Here we used the USPS database for multiclass classification. This database contains 7,291 training patterns and 2,007 testing patterns of 16 × 16 images.[21] We selected 500 samples each digit for training and all testing patterns for test. The original USPS data are saved using $[-1, +1]$ range to represent patterns and we normalized the data to the range $[0, 1]$.

Again, we adopted nonnegative matrix factorization[10] to get sparse features with lower dimensionality by setting $d = 100$ simply and heuristically. We set four values for λ for comparison: 0, 0.1, 0.2, and 0.3. We considered two cases with different mixture components for each class: $M = 24, 48$. All models were initialized randomly where the dimension of each Θ_j is equal to d, and trained by the same 100 iterations to converge for comparison.

It can be seen from Table 2 that the accuracy of classification increases with more mixture components (M) per class. To some extent, the proposed approach with positive value of λ can improve the generalization ability in contrast to the original method, although it is marginally improved. One reason for this may be lack of enough training samples, and the other comes from the limitation of the method, i.e. how to select optimal λ and M for the best generalization as stated above.

Table 2. Classification accuracy rates (%) on USPS data set where **TR** stands for training data set and **TE** for test.

	$\lambda = 0$		$\lambda = 0.1$		$\lambda = 0.2$		$\lambda = 0.3$	
M	TR	TE	TR	TE	TR	TE	TR	TE
24	88.74	84.55	88.78	84.55	88.80	84.60	88.76	84.60
48	95.20	88.99	95.30	89.14	95.30	89.04	95.26	89.19

6. Conclusion and Future Work

Starting with analyzing the discriminative objective function, we show its relation to maximum likelihood and maximum entropy principles and that poor generalization occurs without *prior* knowledge. This paper investigates a sparse mixture model for discriminative classification via nonnegative matrix factorization in order to improve the generalization. The basic idea of our method is to minimize the sum of the weights of un-normalized mixture models for *posterior* according to regularization theory. The experiments on CBCL face database for binary classification and on USPS digit data set for multiclass classification assess the validity of the proposed method.

It is easy to use the proposed approach for other classification tasks, such as speech recognition,[20] and compare this method to others such as support vector machines,[22] and further apply it for face detection.[24] These are future works for us to consider. In addition, we think it is important to investigate how to decide the optimal regularization parameter to get the best generalization. And finally it is worthy of considering different basis functions for mixture models, such as Gaussian mixtures.[7]

Appendix A

Proof of convergence

For simplicity and convenience, we follow the steps in Ref. 19. Adopt the same notations, first let

$$P^{+}_{nij} = \frac{Z_{ni} W_{ij} e^{\Theta_l \cdot X_n}}{\sum_l W_{il} e^{\Theta_j \cdot X_n}}, \quad P^{-}_{nij} = \frac{W_{ij} e^{\Theta_j \cdot X_n}}{\sum_{kl} W_{kl} e^{\Theta_l \cdot X_n}}. \tag{21}$$

Then we recall three Eqs. (15)–(17) from Ref. 19, as below

$$L'_{+} = \sum_n \log \sum_{ij} Z_{ni} W'_{ij} e^{\Theta_j{}' \cdot X_n} \geq \sum_{nij} P^{+}_{nij} \log \left[\frac{Z_{ni} W'_{ij} e^{\Theta'_j \cdot X_n}}{P^{+}_{nij}} \right], \tag{22}$$

$$L'_{-} - L_{-} = \sum_n \log \left(\frac{\sum_{ij} W'_{ij} e^{\Theta'_j \cdot X_n}}{\sum_{kl} W_{kl} e^{\Theta_l \cdot X_n}} \right) \leq \sum_n \left(\frac{\sum_{ij} W'_{ij} e^{\Theta'_j \cdot X_n}}{\sum_{kl} W_{kl} e^{\Theta_l \cdot X_n}} - 1 \right), \tag{23}$$

and

$$e^{\Theta'_j \cdot X_n} \leq e^{\Theta_j \cdot X_n} + \sum_u \left(e^{s\Theta'_{ju}} - e^{s\Theta_{ju}} \right) \left(\frac{X_{nu} e^{\Theta_j \cdot X_n}}{s e^{s\Theta_{ju}}} \right). \tag{24}$$

Now we can get the following inequality

$$\begin{aligned} L_r' - L_r \geq & \sum_{nij} P_{nij}^{+} \left[\log \left(\frac{W_{ij}'}{W_{ij}} \right) + (\Theta_j' - \Theta_j) \cdot X_n \right] \\ & - \sum_{nij} P_{nij}^{-} \left[\frac{W_{ij}'}{W_{ij}} - 1 + \frac{W_{ij}'}{W_{ij}} \sum_u X_{nu} \left(\frac{e^{s(\Theta_{ju}' - \Theta_{ju})} - 1}{s} \right) \right] \\ & - \lambda \sum_{ij} (W_{ij}' - W_{ij}). \end{aligned} \tag{25}$$

Both sides of the above inequality become zeros when $W_{ij}' = W_{ij}$ and $\Theta_j' = \Theta_j$. Finally, maximize the right-hand side w.r.t. W_{ij}' with the basis function parameters fixed and one can get the update rule (16); maximize the right-hand side w.r.t. Θ_j' with the mixture weights fixed and one can get the update rule (17). All update rules keep the same property as those in Ref. 19.

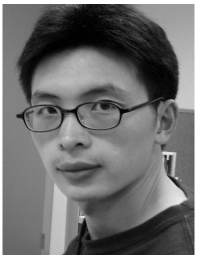
Acknowledgments

The authors acknowledge and thank the creators of both USPS digit database and CBCL face data set for making their data sets publicly available. The authors are grateful to Zhe Chen for sending his paper and good ideas on some topics related to the regularization method. The authors would also like to thank all the anonymous reviewers for their comments to improve the paper. This work was supported by the National Natural Science Foundation of China (Grant No. 60021302).

References

1. CBCL Face Database #1, MIT Center For Biological and Computation Learning, http://www.ai.mit.edu/projects/cbcl.
2. Z. Chen and S. Haykin, On different facets of regularization theory, *Neural Comput.* **14** (2002) 2791–2846.
3. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd edn. (John Wiley & Sons, 2001).
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edn. (Morgan Kaufmann, 1990).
5. P. O. Hoyer, Non-negative sparse coding, *Neural Networks for Signal Processing XII, Proc. IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland (2002), pp. 557–565.
6. T. S. Jaakkola and D. Haussler, *Exploiting Generative Models in Discriminative Classiers*, Advances in Neural Information Processing Systems (1999), pp. 487–493.
7. A. Klautau, N. Jevtic and A. Orlitsky, Discriminative Gaussian mixture models: a comparison with Kernel classifiers, *Proc. Twentieth Int. Conf. Machine Learning (ICML-2003)*, Washington DC (2003).
8. D. Keysers, F. J. Och and H. Ney, Maximum entropy and Gaussian models for image object recognition, *DAGM 2002, Pattern Recognition, 24th DAGM Symp.* Zürich, Switzerland, Lecture Notes in Computer Science, Vol. 2449 (Springer-Verlag, 2002), pp. 498–506.

9. D. D. Lee and L. K. Saul, *Algorithms for Nonnegative Matrix Factorization*, Advances in Neural Information Processing Systems (2001), pp. 556–562.
10. D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* **401** (1999) 788–791.
11. S. Z. Li, X. W. Hou and H. J. Zhang, Learning spatially localized parts-based representation, *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition* (2001).
12. W. Liu, N. Zheng and X. Lu, Non-negative matrix factorization for visual coding, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (2003) III_293–III_296.
13. D. J. C. Mackay, Bayesian methods for adaptive models, Ph.D. thesis, California Institute of Technology, USA (1991).
14. J. B. Paris, Common sense and maximum entropy, *Synthese* **117** (1999) 75–93.
15. G. Rätsch, T. Onoda and K.-R. Müller, Soft margins for AdaBoost, *Mach. Learn.* **42**(3) (2001) 287–320.
16. S. Raudys, *Statistical and Neural Classifiers: An Integrated Approach to Design* (Springer, London, 2001).
17. Y. D. Rubinstein and T. Hastie, Discriminative vs informative learning, *Proc. Third Int. Conf. Knowledge Discovery and Data Mining* (1997), pp. 49–53.
18. W. S. Sarle, Neural Network FAQ, Part 3 of 7: Generalization, *Periodic Posting to the Usenet Newsgroup Comp.ai.neural-nets*, ftp://ftp.sas.com/pub/neural/FAQ.html (2001).
19. L. K. Saul and D. D. Lee, *Multiplicative Updates for Classification by Mixture Models*, Advances in Neural Information Processing Systems (2002).
20. R. Schlüter, B. Müller and H. Ney, Comparison of discriminative training criteria and optimization methods for speech recognition, *Speech Commun.* **34** (2001) 287–310.
21. USPS data set, available on http://www.kernel-machines.org/data.html.
22. V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn. (Springer, NY, 2000).
23. P. M. Williams, Bayesian regularisation and pruning using a Laplace prior, *Neural Comput.* **1** (1994) 425–464.
24. M. H. Yang, D. Kriegman and N. Ahuja, Detecting faces in images: a survey, *IEEE Trans. Patt. Anal. Mach. Intell.* **24** (2002) 34–58.
25. T. Zhang, *Regularized Winnow Methods*, Advances in Neural Information Processing Systems (2001), pp. 703–709.



Weixiang Liu received the B.Sc. and M.Sc. degrees in mechanical engineering from Xi'an Shiyou University, China, in 1997 and 2000 respectively, and the Ph.D. in electronic & information engineering, at the Institute of Artificial Intelligence and Robotics, from Xi'an Jiaotong University, China in 2005.

His research interests include computer vision and pattern recognition, machine learning and information geometry.

Songfeng Zheng received his B.E. degree from the Department of Information and Communication Engineering, and his M.S. degree from the Department of Computer Science, in 2000 and 2003, respectively; both degrees are from Xian JiaoTong Univerisity, China. He is currently a Ph.D. student in the Statistics Department at the University of California, Los Angeles.

His research interests include computer vision, machine learning, pattern recognition and statistical modeling.

Nanning Zheng graduated in 1975 from the Department of Electrical Engineering, Xi'an Jiaotong University, China, and received the M.E. degree in information and control engineering from Xi'an Jiaotong University, China in 1981, and the Ph.D. in electrical engineering from Keio University, Japan, in 1985. He is currently a professor and the director of the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University.

He served as the general chair for the International Symposium on Information Theory and Its Applications in 2002, and the general co-chair for the International Symposium on Nonlinear Theory and Its Applications in 2002. Since 2000, he has been China representative of the Governing Board of the International Association for Pattern Recognition. He presently serves as executive editor of Chinese Science Bulletin. He became a member of the Chinese Academy Engineering in 1999. He is a Fellow of IEEE.

His research interests include computer vision, pattern recognition, computational intelligence, image processing and hardware implementation of intelligent systems.