Generalized Eigenvectors for Large Multiclass Problems

Luke Vilnis
School of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
luke@cs.umass.edu

Nikos Karampatziakis, Paul Mineiro
Microsoft CISL
1 Microsoft Way
Redmond, WA 98052
{nikosk,pmineiro}@microsoft.com

Abstract

Generalized eigenvectors (GEVs) between the class-conditional second moment matrices have recently been shown to be a scalable and powerful technique for feature learning, demonstrating state-of-the-art performance on a variety of domains. However, the number of such features naturally scales with the square of the number of classes k^2, causing both statistical and computational inefficiencies when k^2 is large. This is a serious weakness of the basic GEV method in light of ongoing interest in large-cardinality multiclass classification. We investigate several methods for feature selection by picking class pairs, and demonstrate the superiority of a novel approach that casts the problem as an adversarial saddle-point problem, solvable with an efficient convex program.

1 Introduction

Convex approaches based on linear models have enjoyed great success, especially for problems involving high-dimensional, sparse binary datasets [1, 2]. Nonconvex (but tractable) bilinear models have also been used for large-cardinality output problems such as retrieval [3, 4] and recommendation systems [5]. For lower-dimensional dense datasets, such as those found in vision and speech processing, there has been a renewed interest in joint learning of the predictor and the representation (often in several layers). These techniques as implemented in *deep neural networks* have seen state of the art performance in vision and speech [6, 7], including classification problems with very large numbers of classes, such as object detection in ImageNet 22k [8].

Joint learning of the classifier and latent representation leads naturally to a nonconvex objective function, and direct optimization via gradient-based methods (such as backpropagation in a deep neural network) can be difficult to tune and run into trouble with local minima and parallelization. To avoid these robustness problems, generalized eigenvalue problems between pairs of class-conditional second moment matrices have recently been introduced as an attractive solution to the problem of discriminative feature learning [9]. GEV feature learning optimizes a non-convex objective, but uses linear algebraic techniques to avoid local minima and enable distribution.

In [9] the authors propose an all-class-pairs feature-learning stage followed by a standard linear multiclass classifier. While the running time for the feature learning stage is independent of the number of examples, it scales with the square of the number of classes k^2, which can make them unsuitable for problems with large output spaces. In this paper, we develop a new method for choosing class pairs for GEV features and demonstrate an improvement over strong baselines. Interestingly, we also note good performance on the 1000-class ALOI [10] when using a number of pairs that is even less than the number of classes k, indicating that the discriminative pair selection can uncover features that generalize across classes.

2 GEV Features

Generalized eigenvector features look at local maximizers v of the quotient

$$R_{ij}(v) = \frac{\mathbb{E}[(v^\top x)^2 | y = i]}{\mathbb{E}[(v^\top x)^2 | y = j]} = \frac{v^\top C_i v}{v^\top C_j v} \tag{1}$$

where C_i and C_j are the covariance (or second moment) matrices conditioned on the ith or jth class. That is, they find directions that capture most of the variance (or second moment) in one class i while capturing very little in another class j, by solving the generalized eigenvalue problem $C_i v = \lambda C_j v$. The eigenvectors are local maximizers of (1) and the eigenvalues are the values of the objective.

For a space of dimension d, this allows us to extract up to d C_j-orthogonal features for each pair, and to assess the feature's discriminative power by examining the eigenvalue λ. These discriminative directions v can then be combined with a simple nonlinearities (such as regression splines $\max(t, v^\top x)$) to provide a set of nonlinear features for the data.

In addition to working well in practice, these features have several desirable theoretical and practical properties. Firstly, they are invariant to invertible linear transformations of the input data, and by working directly with the second moment matrices, we require only an initial map-reduce style pass to aggregate the moments over a large dataset. Finally, they should generalize well if the moments can be correctly estimated, which generally should take $O(d \log d)$ examples per class [11].

3 Methods

We investigate several techniques for picking a parsimonious set of class pairs, including two baselines using randomization and the empirical confusion matrix, and a novel algorithm that that directly optimizes a variant of the original GEV objective function to select pairs.

3.1 Randomization

As introduced in [9] Section 4.3 we can place each class on a vertex of a random hypercube in $\lceil \log k \rceil$ dimensions. For each vertex we solve $\lceil \log k \rceil$ GEV problems between the current vertex and each of the immediate neighbors. The intuition is to create a graph with a single connected component that connects every class to every other class with the fewest number of hops. However, this does not take into account any information about which labels are easily confused with each other. This multiple also leads to $O(k \lceil \log k \rceil)$ pairs being picked, which can be prohibitively large when the number of classes is large, so we also compare to a purely randomized approach on ALOI and MNIST.

3.2 Confusion Matrix

A second natural approach is to use a confusion matrix to pick class pairs. Here we use the confusion matrix from a linear least-squares model on the base feature set to get an asymmetric matrix of costs A. For any given number of pairs N, we can get the top N pairs of classes with the highest costs (excluding the diagonal, of course). The intuition here is that classes that are poorly discriminated by the linear model are in need of nonlinear features that can be provided by the GEVs.

3.3 Adversarial GEV

Since our goal is to pick pairs that are most difficult to discriminate for the features, our approach is to formulate a game for each class in which an adversary attempts to select an opposing class, and we try to select a discriminative vector. Concretely, we formulate this as the following saddle-point problem:

$$\begin{aligned} \min_{\alpha_i} \max_{v} & \frac{v^\top (\sum_{j \neq i} \alpha_{ij} C_j) v}{v^\top C_i v} \\ \text{s.t.} \quad & \sum_j \alpha_{ij} = 1, \;\; \alpha_{ij} \geq 0 \end{aligned}$$

This objective can be solved efficiently and optimally with convex techniques using the bundle method [12]. This amounts to making repeated calls to a vanilla GEV solver while changing mixture weights. This process usually converges with a small number of calls to the GEV solver.

The output of this procedure is a matrix of dual variables α_{ij}, where each row corresponds to a fixed denominator and each column corresponds to the mixture weights chosen by the adversary for that numerator matrix. We can expect this approach to work better than the confusion matrix from a linear model since it directly trades the GEV features off against one another, while the confusion matrix may potentially become "unconfused" in many places by the addition of a single nonlinear feature.

4 Related Work

Many feature extraction and dimensionality reduction techniques in machine learning can be formulated as (generalized) eigenvalue problems, including PCA, Fisher's LDA [13], VCA [14], and variants thereof. A comparison of these methods with class-conditional GEV is given in Karampatziakis and Mineiro [9]. The method is similar to a technique in time-series analysis called *common spatial pattern* (CSP) [15], which finds GEVs between two windows of a multivariate signal.

Large-cardinality multiclass classification is an active research area. Some approaches rely on splitting the set of classes into trees recursively to avoid comparing dissimilar classes [16, 17]. Others embed the labels in a low dimensional space and reason in the continuous space instead of the discrete output set [4]. Other approaches, such as word embeddings trained as classifiers [18], also jointly embed inputs and outputs. This work is most similar to these techniques that co-embed the input and output spaces [18, 4], since it directly extends a base feature learning algorithm to select features that are well suited to the output class distribution. However, our technique does not learn vectors for the labels. While the adversarial technique in this paper is novel, the motivation is to find a tractable relaxation of end-to-end supervised nonconvex feature learning, such as that done by backpropagation [13].

5 Experiments

5.1 MNIST

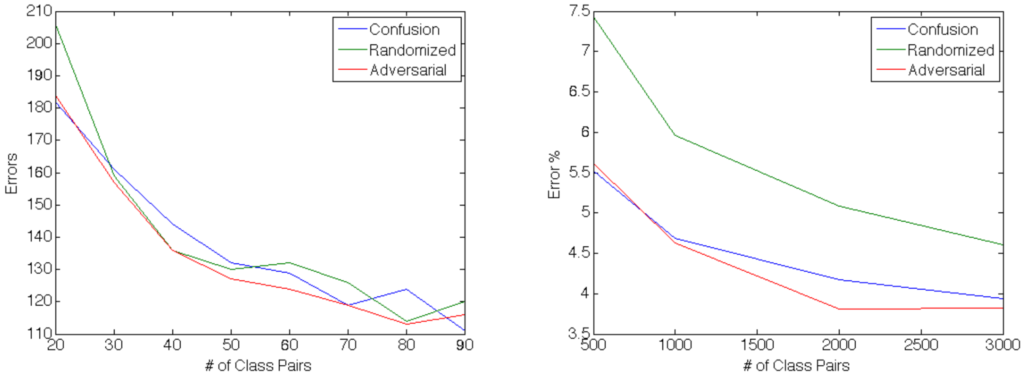
We first compare the different methods for pair-picking on the MNIST dataset of handwritten digits [19], consisting of 10 classes with 784 base pixel features. On this dataset, the number of classes is small enough that the all-pairs approach for generating GEV features is tractable and gives the best performance, so we do not provide a table of results. However, in Figure 1a we can see that error decreases more rapidly when picking pairs using the supervised methods vs. the randomized method, and that the adversarial method mostly dominates the other two.

5.2 TIMIT

We report results on the TIMIT speech-recognition dataset [20] of 183 classes (61 base phonemes split into 3 segment markers), using size-11 windows of standard MFCC features (429 total features). We treat this as a classification problem and score with accuracy, rather than evaluating the whole sequence. We see in Figure 1c that using supervised methods to pick pairs works better than using the randomized hypercube approach. However, in this case using the empirical confusion matrix gives slightly better performance than the adversarial GEVs.

5.3 ALOI

Our largest-cardinality output space comes from the Amsterdam Library of Object Images (ALOI) object classification dataset [10]. This consists of 1000 different classes, using 128 base features derived from color histograms. Here we see that the supervised approaches significantly outperform randomization, and adversarial GEV outperforms the confusion matrix approach for smaller numbers of pairs. This can be expected since the large output space gives appropriate pair-selection an advantage over brute force randomization. The most interesting result here is the good empirical



(a) MNIST: errors vs. number of class pairs.

(b) ALOI: error % vs. number of class pairs.

Model	Error	# Feats
T-DSN [21]	40.9	–
Random GEV [9]	41.87 +/- 0.073	46713
Adversarial GEV	41.78	45009
Confusion GEV	**41.65**	46813

(c) TIMIT: the number of pairs was picked so as to give approximately the same number of features as used in Karampatziakis and Mineiro [9]. Performance of deep neural network T-DSN included for comparison.

Model	Error	# Feats
Linear [17]	13.78	128
RBF [22]	7.0	–
Random GEV (2k)	5.08	4001
Adversarial GEV (2k)	**3.81**	4001
Confusion GEV (2k)	4.17	4001
Random GEV (3k)	4.61	6001
Adversarial GEV (3k)	**3.82**	6001
Confusion GEV (3k)	3.94	6001

(d) ALOI: the number of pairs picked appears in parentheses where appropriate.

Figure 1: Comparison of pair-picking methods for MNIST, TIMIT, and ALOI.

performance even when we use fewer class pairs than the number of classes. This indicates that the GEV features can help discriminate between more than just the two classes for which they were chosen. As seen in Figure 1d, our absolute performance on this dataset is quite good, beating results reported on the same features for the RBF kernel, as well as the linear baseline, by a significant margin.

6 Conclusions and Future Work

We have presented techniques for scaling discriminative GEV features to high-cardinality label space problems. GEV features have several theoretical and practical advantages, including good guarantees on estimation error, scalability, global optimality, and strong empirical performance. Our feature selection method builds on the linear algebraic techniques of the base GEV feature learning algorithm to learn parsimonious sets of features, even with many classes, aiding generalization. We demonstrate the performance of our method compared to natural baselines using randomization and confusion matrices.

There are two clear directions for future work. The first is to test our algorithms on more challenging datasets, such as ImageNet 1k and 22k [8]. The second is to extend our method to incorporate tree structures over the label space, and learn GEVs between clusters of labels. This could aid prediction and estimation for problems where even the sparse pair selection used by our method is intractable.

References

[1] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view

from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230. ACM, 2013.

[2] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *arXiv preprint arXiv:1110.4198*, 2011.

[3] Michael W Berry, Susan T Dumais, and Gavin W O'Brien. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595, 1995.

[4] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation.

[5] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[9] Nikos Karampatziakis and Paul Mineiro. Discriminative features via generalized eigenvectors. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 494–502, 2014. URL `http://jmlr.org/proceedings/papers/v32/karampatziakis14.html`.

[10] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

[11] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[12] Quoc V Le, Alex J Smola, and Svn Vishwanathan. Bundle methods for machine learning. In *Advances in neural information processing systems*, pages 1377–1384, 2007.

[13] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020, 9780262018029.

[14] Roi Livni, David Lehavi, Sagi Schein, Hila Nachliely, Shai Shalev-Shwartz, and Amir Globerson. Vanishing component analysis. In *Proceedings of The 30th International Conference on Machine Learning*, pages 597–605, 2013.

[15] ZoltanJ. Koles, MichaelS. Lazar, and StevenZ. Zhou. Spatial patterns underlying population differences in the background eeg. *Brain Topography*, 2(4):275–284, 1990. ISSN 0896-0267. doi: 10.1007/BF01129656. URL `http://dx.doi.org/10.1007/BF01129656`.

[16] Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In *Advances in Neural Information Processing Systems*, pages 163–171, 2010.

[17] Anna Choromanska and John Langford. Logarithmic time online multiclass prediction. *arXiv preprint arXiv:1406.1822*, 2014.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] W. Fisher, G. Doddington, and Goudie K. Marshall. The DARPA speech recognition research database: Specification and status. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 93–100, 1986.

[21] Brian Hutchinson, Li Deng, and Dong Yu. Tensor deep stacking networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1944–1957, 2013.

[22] Anderson Rocha and Siome Klein Goldenstein. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches.