CLASSIFICATION VIA GROUP SPARSITY PROMOTING REGULARIZATION

A. Majumdar and R. K. Ward

Department of Electrical and Computer Engineering, University of British Columbia
{angshulm, rababw}@ece.ubc.ca

ABSTRACT

Recently a new classification assumption was proposed in [1]. It assumed that the training samples of a particular class approximately form a linear basis for any test sample belonging to that class. The classification algorithm in [1] was based on the idea that **all** the correlated training samples belonging to the correct class are used to represent the test sample. The Lasso regularization was proposed to select the representative training samples from the entire training set (consisting of all the training samples). Lasso however tends to select a single sample from a group of correlated training samples and thus does not promote the representation of the test sample in terms of **all** the training samples from the correct group. To overcome this problem, we propose two alternate regularization methods, Elastic Net and Sum-Over-l_2-norm. Both these regularization methods favor the selection of multiple correlated training samples to represent the test sample. Experimental results on benchmark datasets show that our regularization methods give better recognition results compared to [1].

Index Terms— Classification, Face Recognition, Elastic Net, Group Sparse Regularization

1. INTRODUCTION

Recently a new classifier was proposed in [1]. The work makes a novel classification assumption. It assumes that the training samples of a particular class approximately form a linear basis for a new test sample belonging to the same class. The classification algorithm built upon this assumption gave good recognition results on the Extended Yale B face recognition database [2].

We can write the aforesaid assumption formally. If $v_{k,test}$ is the test sample belonging to the k^{th} class then,

$$v_{k,test} = \alpha_{k,1}v_{k,1} + \alpha_{k,2}v_{k,2} + ... + \alpha_{k,n_k}v_{k,n_k} + \varepsilon \quad (1)$$

where $v_{k,i}$ are the training samples and ε is the approximation error.

In a classification problem, the training samples and their class labels are provided. The task is to assign the given test sample with the correct class label. This requires finding the coefficients $\alpha_{k,i}$ in equation (1). In [1] the solution is framed as a sparse optimization problem. In this work we propose alternate solutions and show that our solutions give better results compared to the previous one [1].

Equation (1) expresses the assumption in terms of the training samples of a **single** class. Alternately, it can be expressed in terms of **all** the training samples so that

$$v_{k,test} = V\alpha + \varepsilon \quad (2)$$

where $V = [v_{1,1} | ... | v_{n,1} | ... | v_{k,1} | ... | v_{k,n_k} | ...v_{C,1} | ... | v_{C,n_C}]$

and $\alpha = [\alpha_{1,1}...\alpha_{1,n_1}...\alpha_{k,1}...\alpha_{k,n_k}...\alpha_{C,1}...\alpha_{C,n_C}]'$.

There are two implications that follow from the assumption expressed in equation (2):

1. The vector $\boldsymbol{\alpha}$ should be sparse.
2. **All** (or most of) the training samples corresponding to the correct class should have non-zero values in $\boldsymbol{\alpha}$.

The above implications demand that $\boldsymbol{\alpha}$ should be 'group sparse' - meaning that the solution of the inverse problem (2) should have non-zero coefficients corresponding to a particular group of correlated training samples and zero elsewhere. The solution in [1] is based on the first implication only. It imposes Lasso regularization on equation (2). Lasso promotes a sparse solution of $\boldsymbol{\alpha}$ but does not favor grouping of correlated samples. Consequently the non-zero values in $\boldsymbol{\alpha}$ do not necessarily correspond to training samples belonging to the same group. Our work proposes two alternate regularizations that promote group sparsity in $\boldsymbol{\alpha}$. Experimental evaluation shows that our method provides better recognition results compared to [1].

The rest of the paper will be organized into several sections. In section 2, we will discuss the background of the problem. Section 3 will detail our proposed methods. In section 4 we will show the experimental results. Finally in section 5, conclusions and future scope of work will be discussed.

2. REVIEW OF PREVIOUS WORK

The novel classification assumption was first proposed in [1]. The first step towards classification is to solve for the coefficient vector $\boldsymbol{\alpha}$ in equation (2). The simplest solution to equation (2) involves the pseudo-inverse of V and is expressed as $\hat{\alpha} = (V'V)^{-1}V'v_{k,test}$. However, in most cases

the matrix V is ill-conditioned or ill-posed. So the simple solution involving the pseudo-inverse is not stable.

To obtain a stable solution, one requires regularizing equation (2) in order to find an approximate stable solution. Since the solution $\hat{\alpha}$ should be sparse, an l_0-norm regularizer is required and the following optimization problem needs to be solved

$$\min_{\alpha} \| \alpha \|_0 \text{ such that } \| v_{k,test} - V\alpha \|_2 < \varepsilon \quad (3)$$

In [1], it is argued that solving the l_0-norm is an NP hard problem and there is no tractable algorithm to solve it. Citing studies in Compressive Sampling [3], they argued that for large systems the l_0-norm can be replaced by the l_1-norm (Lasso regularization) which also leads to a sparse solution.

$$\min_{\alpha} \| \alpha \|_1 \text{ such that } \| v_{k,test} - V\alpha \|_2 < \varepsilon \quad (4)$$

The optimization problem in equation (4) can be solved by quadratic programming methods.

Once a sparse solution of $\boldsymbol{\alpha}$ is obtained, the following classification algorithm was proposed to determine the class of the test sample.

Algorithm 1

1. Solve the optimization problem expressed in (4).
2. For each class i repeat the following two steps:
   a. Reconstruct a sample for each class by a linear combination of the training samples belonging to that class by the equation
   $$v_{recon}(i) = \sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j} \ .$$
   b. Find the error between the reconstructed sample and the given test sample by
   $$error(v_{test}, i) = \| v_{k,test} - v_{recon(i)} \|_2$$
3. Once the error for every class is obtained, choose the class having the minimum error as the class of the given test sample.

l_1-norm minimization leads to a sparse solution, but there are spurious coefficients in the vector $\boldsymbol{\alpha}$ associated with the samples that do not belong to the class of the test sample. Step 2 is required to eliminate the effects of these coefficients. The coefficients in $\alpha_{i,j}$ corresponding to each class i are used to reconstruct a sample. For each class the error between the reconstructed sample and the given test sample is calculated. The assumption in equation (1) says that the error between the reconstruction and the test sample will be the least for the correct class. Based on this assumption, the identity of the test sample is decided by the minimum error.

3. PROPOSED CLASSIFICATION METHODS

We mentioned in section 1 that the assumption in [1] leads to two implications. The previous work [1] based their solution on the first implication, i.e. on the sparsity of the solution. It did not account for the group sparsity of $\boldsymbol{\alpha}$. In this section we will introduce regularizations that make $\boldsymbol{\alpha}$ group sparse, i.e. it has non-zero coefficients corresponding to a particular group of correlated training samples and zero elsewhere.

3.1. Disadvantage of Lasso Regularization

There is a limitation to the Lasso (l_1-norm) regularization. If there is a group of samples whose pair-wise correlations are very high, then Lasso tends to select one sample only from the group [5].

In a classification problem, training samples belonging to the same class are correlated with each other. In such a situation the Lasso regularization proposed in [1] tends to select only a single training sample from the entire class. Thus, in the extreme case, the classifier in [1] becomes a scaled version of the Nearest Neighbour (NN) classifier.

For explaining this effect of the Lasso regularization we rewrite the assumption expressed in equation (1):

$$v_{k,test} = \alpha_{k,1} v_{k,1} + \alpha_{k,2} v_{k,2} + ... + \alpha_{k,n_k} v_{k,n_k} + \varepsilon$$

where the v_k's belong to the same class and are correlated with each other. If algorithm 1 is employed for classifying the test sample, then (in the extreme case) we find that

1. The Lasso regularization tends to select only one of the training samples from the group. We call it $v_{k,best}$.
2. Step 2 is repeated for each class.
   a. The reconstructed vector becomes a scaled version of the selected sample, i.e. $v_{recon}(i) = \alpha_{i,best} v_{i,best}$.
   b. The error from the reconstructed vector is calculated
   $$error(v_{test}, i) = \| v_{k,test} - \alpha_{i,best} v_{i,best} \|_2 \ .$$
3. The class with the minimum error is assumed to be the class of the test sample.

The minimum Lasso error in step 2.b is $\| v_{k,test} - \alpha_{k,best}^{Lasso} v_{k,best}^{Lasso} \|_2$. In NN classification the criterion for choosing the class of the test sample is $\| v_{k,test} - v_{i,j} \|_2 \ \forall \text{ j} \in \text{class i}$. This error is minimized when $v_{i,j} = v_{k,best}^{NN}$ and is given by $\| v_{k,test} - v_{k,best}^{NN} \|_2$. The Lasso error and the NN error are the same except for the scaling factor ($\alpha_{k,best}^{Lasso}$).

When the training samples are highly correlated (which generally is the case in classification), employing Lasso regularization forms a serious limitation to the sparse classification problem. Decision regarding the correct class of the test sample should depend on all the training samples belonging to a class. But Lasso favors selecting a single training sample. We look for alternate regularization methods to overcome this problem.

3.2. Elastic Net Regularization

The problem of selecting a sparse group is studied in [5, 6] where an alternate regularization called 'Elastic Net' that promotes selection of sparse groups is proposed. We apply this regularization to the classification problem.

We repeat the optimization problem used in [1]

$$\min_{\alpha} \| \alpha \|_1 \text{ such that } \| v_{k,test} - V\alpha \|_2 < \varepsilon$$

This has the equivalent (Lasso) expression

$$\min_{\alpha} \| v_{k,test} - V\alpha \|_2 \text{ such that } \| \alpha \|_1 < \tau$$

The unconstrained form of Lasso regularization is

$$\min_{\alpha} \| v_{k,test} - V\alpha \|_2 + \lambda \| \alpha \|_1 \quad (5)$$

To promote group sparsity, Elastic Net regularization, proposes the following optimization problem

$$\min_{\alpha} \| v_{k,test} - V\alpha \|_2 + \lambda_1 \| \alpha \|_2^2 + \lambda_2 \| \alpha \|_1 \quad (6)$$

The l_1 penalty in the above expression promotes sparsity of the coefficient vector $\boldsymbol{\alpha}$, while the quadratic l_2 penalty encourages grouping effect, i.e. selection of a group of correlated training samples. The combined effect of the mixed penalty term is that it enforces group-sparsity, i.e. the recovery of one or very few groups of correlated samples.

The classification is performed by algorithm 1, but instead of solving the optimization problem in equation (4) we need to solve the problem in equation (6). The Elastic Net regularization problem was solved using the 'elasticnet' package [7].

3.3. Sum-Over-l_2-norm Regularization

In section 1, we mentioned two implications of the assumption expressed in equation (1). The Lasso exploits only the first implication while Elastic Net exploits both. The Elastic Net regularization is better than the Lasso in the sense that it promotes the selection of one or very few groups of samples. Elastic Net regularization however, does not exploit the labels of the training samples (columns of V). When the labels are known a stronger group sparsity constraint than the Elastic Net can be imposed.

When the column labels of the matrix V is known, a stronger group sparsity promoting regularization [8, 9] can be employed

$$\min_{\alpha} \| \mathrm{A}_1 \|_2 + \| \mathrm{A}_2 \|_2 + ... + \| \mathrm{A}_C \|_2$$

$$\text{such that } \| v_{k,test} - V\alpha \|_2 < \varepsilon \quad (7)$$

$$\text{where } \mathrm{A}_i = [\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n_i}], \text{ for i = 1,2,...,C}$$

The formulation is similar to the Elastic Net regularization. The l_2-norm over the group of correlated variables (A_i's) enforces the selection of the entire group of samples whereas the summation over the l_2-norm ($\sum \mathrm{A}_i$) enforces group sparsity, i.e. the selection of one or very few classes.

The optimization problem (7) requires the label of each column in the matrix V, i.e. the class the column belongs to. In classification tasks, the labels of the training samples are always available, and hence we can use the Sum-Over-l_2-norm regularization (7) for our problem. We propose a slightly modified version of algorithm 1 in this case.

Algorithm 2

1. Solve the optimization problem expressed in (7).
2. Find those i's for which $\|\mathrm{A}_i\|_2 > 0$.
3. For those classes i satisfying the condition in step 2, repeat the following two steps:
   a. Reconstruct a sample for each class by a linear combination of the training samples in that class via the equation $v_{recon}(i) = \sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j}$.
   b. Find the error between the reconstructed sample and the given test sample by $error(v_{test}, i) = \| v_{k,test} - v_{recon(i)} \|_2$
4. Once the $error(v_{test}, i)$ for every class i is obtained, choose the class having the minimum error as the class of the given test sample.

The computational cost of algorithm 2 is less than algorithm 1, because step 3 is not repeated for all the classes. Instead we evaluate only those classes for which there are non-zero entries in the coefficient vector α (step 2).

4. EXPERIMENTAL RESULTS

We performed two sets of experiments. In the first set we apply the sparse classification algorithms on some benchmark databases from the University of California Irvine Machine Learning (UCI ML) repository [8]. Databases that do not have missing values in feature vectors or unlabeled training data were chosen.

In the second set of experiments, we compared the recognition accuracy on the different sparse classifiers for the face recognition task on the Extended Yale B face database. The sparse classification algorithm [1] was originally proposed to address the face recognition problem.

Table 1, shows the classification results on the UCI ML databases. The results are obtained by Leave-One-Out validation. We compare the classification algorithm in [1] against ours. We use the Nearest Neighbour (NN) classifier as a benchmark.

Table 1: Recognition Accuracy of different methods

Name of Dataset	Recognition Accuracy (%)			
	Lasso [1]	Elastic Net	Sum-Over-l_2-norm	NN
Page Block	94.78	95.32	**95.66**	93.34

Abalone	**27.17**	**27.17**	**27.17**	26.67
Segmentation	**96.31**	94.09	94.09	**96.31**
Yeast	57.75	58.23	**58.94**	57.71
German Credit	69.32	72.67	**74.54**	**74.54**
Tic-Tac-Toe	78.89	**84.41**	**84.41**	83.28
Vehicle	65.58	72.34	**73.86**	**73.86**
Australian Cr.	85.94	85.94	**86.66**	**86.66**
Balance Scale	93.33	94.57	**95.08**	93.33
Ionosphere	86.94	**90.32**	**90.32**	**90.32**
Liver	66.68	69.04	**70.21**	69.04
Ecoli	81.53	82.06	**82.88**	80.98
Glass	68.43	69.11	**70.19**	68.43
Wine	**85.62**	**85.62**	**85.62**	82.21
Iris	**96.00**	**96.00**	**96.00**	**96.00**
Lymphography	85.81	86.04	**86.42**	85.32
Hayes Roth	40.23	**41.01**	**41.01**	33.33
Satellite	80.30	80.30	**82.37**	77.00
Haberman	40.52	43.28	43.28	**57.40**

In Table 1, the best results for each dataset are highlighted in bold. Experiments were run on 19 datasets. Our proposed Sum-Over-l_2-norm regularization gave the best results 17 times. Results from our Elastic Net regularization closely followed our Sum-Over-l_2-norm regularization. The recognition results from the Lasso regularization [1] were better than our methods for one case (Segmentation).

For the face recognition experiments, we repeat the experimental set-up in [1]. The experiments are carried on the Extended Yale B Face Database. For each subject, we randomly select half of the images for training and the other half for testing. Table 2 contains the results for face recognition. The features are selected using the Eigenface method. To compare our results with [1], we select the same number of Eigenfaces as proposed in [1].

Table 2: Recognition Accuracies on Extended Yale B

Method	Number of Eigenfaces			
	30	56	120	504
Lasso [1]	86.49	91.71	93.87	96.77
Elastic Net	86.96	92.05	94.26	97.13
Sum-Over-l_2-norm	**89.40**	**93.37**	**95.14**	**97.79**
NN	74.48	81.85	86.08	89.47

The best recognition results are highlighted in bold. It is seen from Table 2 that our proposed Sum-Over-l_2-norm regularization gives the best recognition results for any number of Eigenfaces selected. The Elastic Net regularization is little lower than the Sum-Over-l_2-norm regularization but better than the Lasso.

5. CONCLUSION

A novel classification assumption: states that the training samples of a class approximately form a linear basis for any new test sample" was proposed in [1]. Based on this assumption, a classifier using Lasso regularization was built. We argued that the Lasso regularization is not an ideal choice for the classifier based on the aforesaid assumption as it selects one training sample only to form the basis. To select a basis with many training samples ,we proposed two alternate regularizations techniques, Elastic Net and Sum-Over-l_2-norm for selecting a group of samples. Results on 20 different datasets (19 from the UCI ML repository and Yale Face Database) show that the sparse classifier based on our alternate regularizations yield better recognition results.

The previous work [1] used the sparse classifier for face recognition only. We however, show that the sparse classifier can be used for general purpose classification tasks including face recognition. Using many benchmark datasets from the UCI ML repository, our proposed sparse classifiers is shown to consistently outperform the classifier in [1] and also the NN classifier (see Table 1). For face recognition tasks, our proposed methods, on average, yield around 2% and 11% better recognition accuracy than the classifier in [1] and NN respectively.

6. REFERENCES

[1] Y. Yang, J. Wright, Y. Ma and S. S. Sastry, "Feature Selection in Face Recognition: A Sparse Representation Perspective", IEEE Trans. PAMI, (to appear).

[2] http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

[3] D. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," Comm. on Pure and Applied Math, Vol. 59 (6), pp. 797–829, 2006.

[4] http://www.cs.ubc.ca/labs/scl/spgl1/

[5] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", Journal of Royal Statistical Society B., Vol. 67 (2), pp. 301-320.

[6] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-Net Regularization in Learning Theory", eprint arXiv:0807.3423

[7] http://cran.r-project.org/web/packages/elasticnet/index.html

[8] M. Stojnic, F. Parvaresh and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements", eprint arXiv:0804.0041v1

[9] E. van Den Berg, M. Schmidt, M. P. Friedlander and K. Murphy, "Group Sparsity via Linear-Time Projection", Technical Report TR-2008-09, Department of Computer Science, University of British Columbia

[10] http://archive.ics.uci.edu/ml/