NON-LINEAR SPARSE AND GROUP SPARSE CLASSIFIER

Angshul Majumdar and Rabab K. Ward

Department of Electrical and Computer Engineering, University of British Columbia

ABSTRACT

Recently there has been an interest in a new classification model, where it is assumed that the training samples for a particular class form a linear basis for any new test sample belonging to that class. This assumption led to two successful classification methods called the Sparse Classifier (SC) and the Group Sparse Classifier (GSC). This work generalizes the previous linearity assumption and accounts for non-linear functional relationship between the training samples of a class and a new test sample belonging to that class. Such a generalization requires solving sparse/group–sparse optimization problems with non-linear constraints. We develop exact optimization based algorithms as well as approximate (fast) algorithms to solve such hitherto un-addressed optimization problem. Results show that significant improvement can be achieved by the proposed Non-Linear Sparse Classifiers compared to previous Sparse/Group Sparse Classifiers.

Index Terms— sparsity, group sparsity, non-linear optimization, greedy algorithms.

1. INTRODUCTION

Recently there has been an interest in a new classification model which stems from the assumption that the training samples of a particular class form a linear basis for any new test sample belonging to that class. Expressed formally,

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

where $v_{k,test}$ is the test sample, assumed to belong to k^{th} class, $v_{k,i}, i \in (1, n_k)$ are the training samples for that class and $\alpha_{k,i}$ are the weights.

Equation (1) expresses the assumption in terms of training samples of a single class. It can be expressed in terms of training samples of all the classes,

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

where $V = \{\underbrace{v_{1,1} | ... | v_{1,n_1}}_{v_1}, ..., \underbrace{v_{C,1} | ... | v_{C,n_C}}_{v_C}\}$,

$\alpha = [\underbrace{\alpha_{1,1}, ..., \alpha_{1,n_1}}_{\alpha_1}, ..., \underbrace{\alpha_{C,1}, ..., \alpha_{C,n_C}}_{\alpha_C}]^T$.

According to the assumption, the vector α should be sparse, i.e. zeros everywhere except for weights corresponding to the correct class. The equation (2) is ill-posed, therefore to solve for α ones needs to impose certain sparsity group-sparsity constraints to obtain a stable solution. This leads to the sparse [1] and group-sparse [2] classifiers. We will discuss more about the constraints in the next section.

The main objective of this paper is to generalize the linearity assumption of (1) and (2), i.e. we will assume a non-linear functional relationship between the training samples and the test sample. Mathematically the non-linearity assumption is expressed as,

$$v_{k,test} = f(V\alpha) \quad (3)$$

where f is a non-linear function.

We will impose sparsity/group-sparsity constraints as in [1] and [2] to solve (3). This will lead to the solution of non-linear inverse problem (3) subject to sparsity/group-sparsity constraints.

The extension from linear to non-linear functional relationship on one hand gives a huge flexibility to the classification assumption, but on the other, it poses to be a very challenging problem to solve. Non-linear inverse problems with sparsity/groups-sparsity constraints have not been encountered earlier. For the first time in this work, we provide exact (but slow) and greedy (approximate) algorithms to solve the said problem.

The rest of the paper is organized into several sections. Section 2, briefly discusses the previous sparse and group-sparse classification algorithms. In Section 3, exact and greedy algorithms to solve the proposed non-linear inverse problem will be described. Experimental results are provided in Section 4. Conclusion of this work is discussed in Section 5.

2. LINEAR SPARSE/GROUP-SPARSE CLASSIFIERS

The sparse/group-sparse classification model leads to the inversion of the linear problem (2)

$$v_{k,test} = V\alpha$$

According to the assumption, the vector α has zeros everywhere else apart from the coefficients corresponding to the training samples of the correct class. In other words, the vector α is sparse.

In general the inversion (2) is under-determined and/or ill-conditioned. Therefore, to obtain the desired solution the inverse problem needs to be regularized. In [1], the

978-1-4799-0033-6/13/$31.00 ©2013 IEEE

inversion problem was solved with sparsity constraints, which led to the following optimization problem,

$$\min \| \alpha \|_1 \quad subject\ to\ v_{k,test} = V\alpha \qquad (4)$$

Such inverse problems with sparsity constraints arise in different branches of signal processing and machine learning; therefore there are quite mature algorithms to solve this problem. We are unable to discuss them due to lack of space.

A closer look at the classification assumption shows that the vector α is not just any sparse vector – it has non-zero coefficients only corresponding to training vectors of the correct class. In other words, the coefficient vector is group-sparse. An optimal inversion problem accounting for group-sparsity was proposed in [2],

$$\min \| \alpha \|_{2,1} \quad subject\ to\ v_{k,test} = V\alpha \qquad (5)$$

where $\| \alpha \|_{2,1} = \sum_{i=1}^{C} \| \alpha_i \|_2$.

Group-sparse inversion is not quite as matured as sparse inversion; but there are a handful of papers that proposes powerful algorithms to solve problems like (5). All the group-sparse inversion algorithms are modifications of popular sparse inversion algorithms.

Sparse [1] and Group-Sparse [2] classifiers were based on the optimization problems (4) and (5) respectively. Although such optimization problems are convex and give good results, they are comparatively slow. In order to address the problem of speed, greedy (sub-optimal) approximate algorithms have been employed to solve the aforesaid inversion problems.

Greedy algorithms for sparsity promoting inverse problem is a matured research area. In [3], it was shown that quite good results can be obtained when (4) is replaced by an Orthogonal Matching Pursuit (OMP) or Orthogonal Least Squares (OLS) [4] based greedy algorithm at a fraction of the time required by exact optimization.

Inversion problems with group-sparsity constraint is a new area. Until recently, there had been no greedy algorithm to solve (5). A suite of greedy algorithms for solving (5) was developed by us in a previous work [5]. The sparse and the group-sparse classifiers which were based on greedy algorithms instead of exact optimization have been termed Fast Sparse Classifiers (FSC) and Fast Group Sparse Classifiers (FGSC) respectively.

Till now we have been discussing solely on the inversion problem. Once the inversion problem is solved, the actual classification algorithm proceeds as follows:

1. Solve the inversion problem (4) or (5).
2. For each class (i) repeat the following two steps:

2.1 Reconstruct a sample for each class by a linear combination of the training samples belonging to that class using $v_{recon}(i) = \sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j}$.

2.2 Find the error between the reconstructed sample and the given test sample by $error(v_{k,test}, i) = \| v_{k,test} - v_{recon(i)} \|_2$.

3. Once the error for every class is obtained, choose the class having the minimum error as the class of the given test sample.

2. PROPOSED NON-LINEAR CLASSIFIERS

In this work, we are interested in solving (3) with sparsity/group-sparsity constraints. Therefore the optimization problems were interested to solve are the following,

$$\min \| \alpha \|_1 \quad subject\ to\ v_{k,test} = f(V\alpha) \qquad (6a)$$

$$\min \| \alpha \|_{2,1} \quad subject\ to\ v_{k,test} = f(V\alpha) \qquad (6b)$$

Equation (6a) leads to the Sparse Non-Linear Classifier (SNLC) and (6b) leads to the Group-Sparse Non-Linear Classifier (GSNLC).

Such non-linear inversion problems with sparsity/group-sparsity constraints have not been encountered before. Therefore there are no tailored algorithms to solve them. General purpose convex optimization packages like CVX [6] can be used to solve them, but they will be prohibitively slow. We will develop exact optimization algorithms as well as greedy approximate ones to solve equations (6). But before going into them, we will discuss how the classification proceeds assuming that the inversion problem is solved.

1. Solve the non-linear inversion problem (6).
2. For each class (i) repeat the following two steps:

2.1 Reconstruct a sample for each class by a linear combination of the training samples belonging to that class using $v_{recon}(i) = f(V\alpha^{(i)})$, where $\alpha^{(i)} = \begin{cases} \alpha_i & \text{for the i}^{th} \text{ class} \\ 0 & \text{for other classes} \end{cases}$.

2.2 Find the error between the reconstructed sample and the given test sample by $error(v_{k,test}, i) = \| v_{k,test} - v_{recon(i)} \|_2$.

3. Once the error for every class is obtained, choose the class having the minimum error as the class of the given test sample.

3.1. Exact Algorithm

The problem in equation (6) in general can be written in the following form,

$$\min E(\alpha)\ subject\ to\ v_{k,test} = f(V\alpha) \qquad (7)$$

where $E(x)$ is $||\alpha||_1$ for sparsity constraint, and $||\alpha||_{2,1}$ for group-sparsity constraint.

There is no closed form solution to (6), it needs to be solved iteratively. In each iteration, it is possible to express

$E(\alpha)$ as a weighted l_2-norm. i.e. $E(\alpha) = \| W\alpha \|_2$ for both the sparsity and group-sparsity constraints:

Sparsity: $w_{i,j} = | \alpha_{i,j} |^{\frac{p}{2}-1} + \varepsilon$ (8a)

Group-Sparsity $w_{i,j} = \| \alpha_i \| \cdot | \alpha_{i,j} |^{\frac{p}{2}-1} + \varepsilon$ (8b)

The weight matrix W, is a diagonal matrix formed by the elements $w_{i,j}$. At each iteration, the weights are computed from the previous iterate. It can be easily verified that, as the solution converges, the weighted l_2-norm reaches the actual $E(\alpha)$. Since the solution is sparse/group-sparse most of the elements of the weight matrix are going to be infinity; in order to stop the elements of W from blowing up to infinity, the damping factor ε is added.

Using (8a) and (8b), equations (6a) and (6b) can be expressed in each iteration as,

$$\min \| W\alpha \|_2 \quad subject\ to\ y = f(V\alpha) \tag{9}$$

alternately,

$$\min \| u \|_2 \quad subject\ to\ y = f(VW^{-1}u) \tag{10}$$

where $u = W\alpha$

At each iteration, it is possible to solve (10) by IPOPT [7]. Once, (10) is solved, the value of solution at that iteration can be found from $x = W^{-1}u$. The pseudo-code for the algorithm is as follows:

Initialize: x, ε and W=*Identity*.
At iteration t:

1. Compute W_t using (8a) or (8b) (as the case may be) from the previous value of the solution, i.e. α_{t-1}.
2. Solve (10) to obtain u_t.
3. Find $\alpha_t = W_t^{-1}u_t$.
4. Check if solution has converged; else go to step 1.

The algorithm is simple to implement. It is motivated by the success of Iterated reweighted least squares (IRLS) algorithms for sparse [8] and group-sparse [9] optimization problems with linearity constraints.

3.2. Greedy Algorithm

Greedy algorithm for the sparse optimization problem with non-linear constraints was proposed in [10]. We propose a variant of [10] to solve our sparse/group-sparse optimization problem.

Initialize: The sparse weight vector is initialized to zero, $\alpha = 0$. The set of chosen indices is empty $L^{(0)} = []$.

At Iteration t:

1. The first step computes the gradient of the error at the current coefficient estimate, i.e. $g = \frac{d}{d\alpha} \| v_{k,test} - f(V\alpha) \|_2^2 \ at\ \alpha^{(k-1)}$. The reason, why the l_2-norm of the residual is because, it has been shown mathematically that the residual follows a Gaussian distribution [11].
2. For the sparse optimization problem, the index having the highest gradient magnitude is chosen. $l = \{i : \max | g(i)) | \}$. For the group-sparsity problem the group having the highest average gradient magnitude is chosen, $l = \{group(i) : \max avg | g(i)) | \}$.
3. Add the current indices to the set of already chosen indices, $L^{(t)} = L^{(t-1)} \bigcup l$.
4. The values of the weights at the chosen indices are computed by least squares optimization $\hat{\alpha} = \min \| v_{k,test} - f(V(:L^{(t)})\alpha) \|_2$. This is an easy problem, since the system of equations $f(V(:L^{(t)})\alpha)$ is over-determined.
5. Update the weight estimate $\alpha = \begin{cases} \hat{\alpha} & \text{for indices in } L^{(t)} \\ 0 & \text{for indices not in } L^{(t)} \end{cases}$.

4. EXPERIMENTAL EVALUATION

In this work, we compare the proposed non-linear sparse and group sparse classifiers with the linear sparse and group sparse classifiers. The experimental evaluation are carried on three benchmark image classification datasets – COIL-20 [12], USPS [13] and Yale B [14].

The Columbia Object Image Library (COIL-20) consists of normalized gray-scale images of 20 objects. Each object has 72 images for different orientations. For our purpose, we randomly chose half of the images for each object for training and the other half for testing.

The USPS dataset stands for the United States Postal Service dataset for handwritten digit recognition. The database contains 9298 samples divided into training and testing sets consisting of 7291 and 2007 samples respectively.

The face recognition experiments were carried out on the benchmark Yale B database. Only the frontal faces were chosen for our experiment. The database consists of 2,414 frontal faces under varying illumination condition for 38 individuals. One half of the images (for each individual) were randomly selected for training and the other half for testing.

For all the datasets, Principal Component Analysis was chosen as the preferred feature extraction method. For classification, we have a host of classification models to choose from. In fact, any polynomial can be used as the classification model. However, keeping in mind the Occam's razor, we experiment only with the following few simple models:

$f_1(A,x) = (Ax)^2 + Ax$

$f_2(A,x) = (Ax)^{1/2} + (Ax)^2 + Ax$

$f_3(A,x) = (Ax)^{1/2} + Ax$

In tables 1-3, the classification errors (in percentage) for COIL-20, USPS and Yale B are tabulated respectively. The aforesaid non-linear classification models are compared against the previously proposed linear models. For all the experiments on different databases, the number of principal components was kept at approximately 12% of the original dimensionality of the samples.

Table 1: Classification Error on COIL-20

Classifier	Model	Exact	Greedy
Sparse Classifiers	Linear [1]	0.11	0.12
	$(Ax)^2 + Ax$	0.11	0.12
	$(Ax)^{1/2} + (Ax)^2 + Ax$	**0.09**	**0.11**
	$(Ax)^{1/2} + Ax$	0.12	0.12
Group Sparse Classifiers	Linear [2]	0.09	0.09
	$(Ax)^2 + Ax$	0.09	0.11
	$(Ax)^{1/2} + (Ax)^2 + Ax$	**0.08**	**0.09**
	$(Ax)^{1/2} + Ax$	0.12	0.12

Table 2: Classification Error on USPS

Classifier	Model	Exact	Greedy
Sparse Classifiers	Linear [1]	5.53	5.59
	$(Ax)^2 + Ax$	**4.17**	**4.31**
	$(Ax)^{1/2} + (Ax)^2 + Ax$	6.72	6.79
	$(Ax)^{1/2} + Ax$	7.03	7.25
Group Sparse Classifiers	Linear [2]	5.34	5.34
	$(Ax)^2 + Ax$	**4.01**	**4.12**
	$(Ax)^{1/2} + (Ax)^2 + Ax$	6.43	6.57
	$(Ax)^{1/2} + Ax$	6.88	7.06

Table 1: Classification Error on Yale B

Classifier	Model	Exact	Greedy
Sparse Classifiers	Linear [1]	4.71	4.96
	$(Ax)^2 + Ax$	5.20	5.27
	$(Ax)^{1/2} + (Ax)^2 + Ax$	3.14	3.30
	$(Ax)^{1/2} + Ax$	**3.02**	**3.19**
Group Sparse Classifiers	Linear [2]	4.32	4.94
	$(Ax)^2 + Ax$	5.09	5.17
	$(Ax)^{1/2} + (Ax)^2 + Ax$	3.02	3.19
	$(Ax)^{1/2} + Ax$	2.87	3.02

Tables 1-3 show that significant improvements can be achieved by considering non-linear models instead of the previously used linear models with sparse/group-sparse constraints.

5. CONCLUSION

This work generalizes a recently proposed classification model, where it was assumed that the test samples of a particular class can be expressed as a linear combination of the training samples belonging to that class. This work generalizes the linear relationship (between the test sample and the training samples) to a non-linear one. This generalization leads to a lead hitherto unaddressed problem of sparsity/group-sparsity promoting optimization problem with non-linear equality constraints.

In this work we propose algorithms for the exact optimization problem as well as fast greedy approximate solutions. The non-linearity assumption leads to potentially innumerable classification models. In this work, we have shown that even with some simple polynomial models, significant improvements in classification accuracy can be achieved with benchmark image classifications databases.

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, Vol. 31 (2), pp. 210-227, 2009.

[2] A. Majumdar and R. K. Ward, "Classification via Group Sparsity Promoting Regularization", ICASSP 2009.

[3] A. Majumdar, and R. K. Ward, "Robust Classifiers for Data Reduced via Random Projections", IEEE Trans. Systems, Man and Cybernetics Part:B, (accepted)

[4] T. Blumensath, M. E. Davies; "On the Difference between Orthogonal Matching Pursuit and Orthogonal Least Squares", manuscript 2007.

[5] Majumdar, A. and Ward R. K., "Fast Group Sparse Classifier", IEEE Canadian Journal for Electrical and Computer Engineering (Invited Paper).

[6] http://www.stanford.edu/~boyd/cvx/

[7] https://projects.coin-or.org/Ipopt

[8] R. Chartrand, and W. Yin, "Iteratively reweighted algorithms for compressive sensing," ICASSP 2008.

[9] A. Majumdar and R. Ward, "Non-Convex Group Sparsity: Application to Color Imaging", accepted ICASSP 2010.

[10] T. Blumensath and M. Davies "Gradient Pursuit for Non-Linear Sparse Signal Modelling", EUSIPCO, 2008

[11] D. Donoho, Y. Tsaig, I. Drori, and J. Starck, "Sparse solutions of underdetermined linear equations by stagewise orthogonal matching pursuit," 2006.

[12] http://www1.cs.columbia.edu/CAVE/publications/pdfs/Nene_TR96.pdf

[13] http://www.cenparmi.concordia.ca/~jdong/HeroSvm.html

[14] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html