A Sparse Multi-class Classifier for Biomarker Screening

Tzu-Yu Liu*, Ami Wiesel†, and Alfred O. Hero*‡
*Electrical Engineering and Computer Science Department, University of Michigan, USA
†School of Computer Science and Engineering, the Hebrew University of Jerusalem, Israel
‡Center for Computational Biology and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

***Abstract*—We introduce a group-sparsity penalized multi-class classifier design that is parameterized by a set of biomarker weight vectors that minimize miss-classification probability of error. The optimization is implemented by augmented Lagrangian and variable splitting methods. This results in a classifier that automatically designates the role of each biomarker included in the classifier. Using our convex optimization approach a multi-class classifier with group-sparsity constraints results in significantly improved classifier performance.**

***Index Terms*—Multi-class classification, variable selection, sparsity, dimension reduction, augmented Lagrangian optimization.**

I. Introduction

High-dimensional applications, such as genomics expression analysis, require parsimonious modeling. By pruning the total number of independent variables or features, variable selection is a first step in building parsimonious models. Accurate variable selection avoids the over fitting problem, and provides interpretations of the most relevant variables for a predictive model. It is important to understand which variables are strongly relevant to the classification task, and how their importance depends on time or subject in the population. For example, sparsity penalized lasso techniques [1], [2] provide a computationally tractable way to perform variable selection driven by objective function minimization. Here we introduce an objective function minimization approach for structured variable selection that adds a mixed L_1/L_2 norm sparsity penalty to the multi-class classification objective function.

The paper is organized as follows. We first present the formulation of the optimization problem in the methods section, including the loss function used as surrogates in multi-class classification, the proper regularization that selects variables relevant simultaneously to all classes and data blocks, and followed by a discussion about the general algorithm we propose to solve the optimization. Then we present the performance of the sparse multi-class classifier applied to a H3N2 flu challenge data set in the results section, with the discussion about the advantages of the methods and biological interpretation. The final section concludes this paper.

II. Methods

Suppose we have a dataset with n samples, $\{\mathbf{x}_i, \mathbf{y}_i, s_i\}_{i=1}^n$, in which $\mathbf{x}_i \in R^{rp}$ are the independent variables, $y_i \in \{1, 2, ..., K\}$ is the dependent variable, and $s_i \in \{1, 2, ..., m\}$ represents the additional information about the generating sources. All r, p, K, m are positive integers. For example, in general serially sampled experiments, multiple measurements are collected over time, contributing one p dimensional measurement at each time point, then $\mathbf{x}_i$ becomes a multi-block data with r blocks. The data may have been collected from m different individuals, labeled as s_i. We propose an algorithm for learning the best classifier of the label y_i given the data $\mathbf{x}_i$ in the high-dimensional case where the number m of subjects is much less than the number p of variables, e.g., gene probes on the microarray.

A. Sparse Multi-class Classifier

To generalize the current methods in binary classification [3] to multi-class problems, we define a unified multi-class classifier. We adopt the Support Vector Machine (SVM) approach. The idea of maximizing the margin between two classes can be extended to multi-class problems. There are two common strategies that have been proposed: (1) solving the multi-class problem by a series of binary SVM classifiers [4]–[6]; (2) formulating a single unified multi-class SVM [7]–[12]. The former approach has the advantage of building on the binary SVM framework; the latter is more direct. We propose a unified multiclass classifier with variable selection following the latter approach, which provides us with the basis for doing structured variable selection over classes and references.

When $r = 1$, the problem reduces to the standard multi-class classification problem, whereas when $r > 1$, this is the multi-block multi-class classification problem. One could reduce this multi-class classification problem into pairs of binary classification problem. However, this does not capture multiway correlations between the different classes. The unified K-class classifier uses rp-dimensional hyperplanes to partition the feature space,

$$F = \{f_1, f_2, \cdots, f_K\},\ where\ f_k(\mathbf{x}) = {\mathbf{w}_k}^T\mathbf{x} + b_k$$

and the decision rule is to assign the label that gives the largest score, $\arg\max_k\{f_k(\mathbf{x})\}$. Thus the problem can be formulated as

$$\min_F \frac{1}{n}\sum_{i=1}^n V(F, \mathbf{x}_i) + \lambda R(F) \tag{1}$$

where V denotes the convex loss function to upper bound the 0-1 loss, and R is a regularization function.

Crammer and Singer [10] introduced a generalized notion of the margin for multi-class problems, and suggested solving the above optimization by using the convex loss

$$V(F,\mathbf{x_i}) = [\max_r(1-\delta_{y_i,r} + f_r(\mathbf{x}_i) - f_{y_i}(\mathbf{x}_i))]_+ \tag{2}$$

and the regularization function

$$R(F) = \frac{1}{2}\sum_{k=1}^{K} ||\mathbf{w}_k||_2^2.$$

R can be chosen as a sparsity inducing penalty [13] [14]. We formulate our reference-based classification problem with the same loss function, but with different regularization to induce sparsity in the weights and perform variable selection.

Define $W = [\ \mathbf{w}_1\ \ \mathbf{w}_2\ \ \cdots\ \ \mathbf{w}_K\ ]^T = [\ \mathbf{w}_{(1)}\ \ \mathbf{w}_{(2)}\ \ \cdots\ \ \mathbf{w}_{(rp)}\ ]$, i.e. $\mathbf{w}_{(j)}$ is the jth column in the matrix W, and $\mathbf{w}_k$ represents the kth row. Given $j \in \{1, 2, ...p\}$, any elements in the ith column of W, such that $mod(i,p) = j$, are related to the same variable j, and the elements in the kth row of W are weights assigned to class k. In multi-class classification problems ($r = 1$), in order to ensure that the predictor variables are shared over all classes, the columns of W should satisfy a coupled sparsity condition: the number of non-zero terms should be small. Variable selection under this framework becomes more complicated than that in binary classification, because one would expect that an unrelated variable corresponds to a zero column in W rather than a zero scalar. Wang and Shen extended L_1 SVM to L_1 MSVM by imposing a penalty with $q = 1$ on the coefficients. They solved a problem of the form [15]:

$$\min_{b,W} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i \neq k)[b_k + \mathbf{w}_k{}^T\mathbf{x}_i + 1]_+ + \lambda\sum_{k=1}^{K}\sum_{j=1}^{p}|w_{kj}|.$$

Although L_1 has the advantage of being directly related to *lasso*, it treats all the w_{kj}'s equally, which does not guarantee the variable sharing condition. Zhang et al accounted for variable sharing by treating the coefficients in groups by imposing a L_∞ penalty as follows [16]

$$\min_{b,W} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i \neq k)[b_k + \mathbf{w}_k{}^T\mathbf{x}_i + 1]_+ + \lambda\sum_{j=1}^{p}||\mathbf{w}_{(j)}||_\infty.$$

When there are more than one block, we propose a regularization of the form:

$$\begin{aligned} R(F) &= \textstyle\sum_{j=1}^{p} ||\tilde{\mathbf{w}}_{(j)}||_2 \\ \tilde{\mathbf{w}}_{(j)} &= [\mathbf{w}_{(j)}; \mathbf{w}_{(j+p)}; ...; \mathbf{w}_{(j+(r-1)p)}]. \end{aligned} \tag{3}$$

This ensures that coefficients corresponding to a shared variable over the measurements under r conditions and over all classes are grouped together, as shown in Figure 1. The L_2 norm instead of the L_∞ norm is chosen in our formulation, because the L_∞ ball tends to favor solution with the scaled version of hadamard matrix, whereas L_2 norm penalizes any direction uniformly.

The adaptive lasso [17], [18] can be applied to our multi-block multi-class formulation to reduce over-fitting. Suppose we have an initial estimation F_{init}. Then adaptive lasso refines this estimate by solving

$$\min_F \frac{1}{n}\sum_{i=1}^{n} V(F,\mathbf{x}_i) + \lambda_{adapt}R_{adapt}(F, F_{init}) \tag{4}$$

in which V is defined as the same generalized hinge loss function and

$$R_{adapt}(F, F_{init}) = \sum_{j=1}^{p} \frac{||\tilde{\mathbf{w}}_{(j)}||_2}{||\tilde{\mathbf{w}}_{init,(j)}||_2}.$$

An interesting property of adaptive lasso is that if the coefficients in the initial estimate are equal to zero, then the new estimate will also be zero. In other words, adaptive lasso is a stage-wise screening process that limits the number of features to be less than or equal to the initial stage, avoiding over-estimation problems. The process can be recursively repeated over multiple stages to successively reduce the number of non-zero weights.

B. Algorithmic Implementation

We solve the optimization problem (1) with the combination of loss function (2) and regularization (3) using variable splitting. Variable splitting is a general approach to solving optimization problems of the form [19]:

$$\min_v f_1(v) + f_2(v). \tag{5}$$

Variable splitting replaces the argument v of f_2 by another variable w. Then the variable splitting optimization $\min_{v,w} f_1(v) + f_2(w)$ becomes equivalent to (5) when one enforces a constraint $v = w$. This constrained problem can be solved as an augmented Lagrangian optimization of the form:

$$\min_{v,w} f_1(v) + f_2(w) + \frac{\mu}{2}||v - w||_2^2$$

which suggests the alternating splitting algorithm (Algorithm 1).

Algorithm 1: Alternating Splitting Method

```
Set $t = 0$, choose $\mu > 0$, $v_0$, $w_0$, and $d_0$;
while stopping criterion is not satisfied do
    $v_{t+1} \in \arg\min_v f_1(v) + \frac{\mu}{2}||v - w_t - d_t||_2^2$;
    $w_{t+1} \in \arg\min_w f_2(w) + \frac{\mu}{2}||v_{t+1} - w - d_t||_2^2$;
    $d_{t+1} = d_t - v_{t+1} + w_{t+1}$;
    $t = t + 1$;
```

Applying these ideas to (1), we split W into two parts W and M, constrained such that $M = W$, and the row vectors $\mathbf{m_k}$ of M obey the same structural pattern as the rows of W. This then leads to the optimization problem

$$\min_{W,M} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\sum_{j=1}^{p}||\tilde{\mathbf{m}}_{(j)}||_2, \ \textit{subject to} \tag{6}$$

$$\forall i,\ k\ \ (\mathbf{w}'_{y_i}\mathbf{x}_i+b_{y_i})+\delta_{y_i,k}-(\mathbf{w}'_k\mathbf{x}_i+b_k)\geq 1-\xi_i,\ M=W.$$

The ξ_i's are slack variables that depend on W through the constraint $M = W$. The optimization of (6) is performed by alternating algorithm 2.

Algorithm 2: Sparse Multi-class Classifier implementation using variable splitting

```
1 set $\tau = 0$, choose $\mu > 0$, $M_0$, $W_0$, $D_0$
2 while stopping criterion is not satisfied do
3     $W_{\tau+1} = \arg\min_{W} \frac{1}{n}\sum_{i=1}^{n} \xi_i + \frac{\mu}{2}||W - M_\tau - D_\tau||_F^2$
4     s.t. $\forall i, k\ \ (\mathbf{w}'_{y_i}\mathbf{x}_i + b_{y_i}) + \delta_{y_i,k} - (\mathbf{w}'_k\mathbf{x}_i + b_k) \geq 1 - \xi_i$
5     $M_{\tau+1} = \arg\min_{M} \lambda \sum_{j=1}^{p} ||\tilde{\mathbf{m}}_{(j)}||_2 + \frac{\mu}{2}||W_{\tau+1} - M - D_\tau||_F^2$
6     $D_{\tau+1} = D_\tau - W_{\tau+1} + M_{\tau+1}$
7     $\tau = \tau + 1$
```

In each iteration, optimization over W is exactly a quadratic programming problem and there exits fast algorithm tailored to SVM classification problems. We adopt the sequential dual method [20], [21]. The dual of line 3 in Algorithm 2 can be written as

$$\begin{aligned}&\min_{\boldsymbol{\alpha}} \tfrac{1}{2}\sum_{k=1}^{K}||\mathbf{w}_k||_2^2+\sum_{i=1}^{n}\sum_{k=1}^{K}\alpha_i^k e_i^k\\ &\text{s.t.} \sum_{k=1}^{K}\alpha_i^k=0,\ \alpha_i^k\leq \tfrac{1}{n\mu}\delta_{y_i,k},\ \forall i,k\end{aligned}$$

in which $\mathbf{w}_k = \sum_{i=1}^{n}\alpha_i^k\mathbf{x}_i + \mathbf{m}_{\tau,k} + \mathbf{d}_{\tau,k}$ and $e_i^k = 1 - \delta_{y_i,k}$. Coordinate descent method can be extended to decompose the dual problem into n subproblems, and each problem corresponds to one of the n samples.

$$\begin{aligned}&\min_{\alpha_i^1,\ldots,\alpha_i^K}\sum_{k=1}^{K}\tfrac{1}{2}A(\alpha_i^k)^2+B_k\alpha_i^k\\ &\text{s.t.}\sum_{k=1}^{K}\alpha_i^k=0,\ \alpha_i^k\leq\tfrac{1}{n\mu}\delta_{y_i,k},\ \forall k\end{aligned}$$

where $A = \mathbf{x}_i\mathbf{x}'_i$ and $B = \mathbf{w}'_k\mathbf{x}_i + e_i^k - A\alpha_i^k$. This is the same optimization problem discussed in [20], [21] except the representation of $\mathbf{w}_k$. We can adopt the subproblem solver based on coordinate descent method.

In the second step, M has a close form solution. Let $C = W_{\tau+1} - D_\tau$, then the solution of each concatenated column of M is given as $\tilde{\mathbf{m}}_{(j)} = [||\tilde{\mathbf{c}}_{(j)}||_2 - \frac{\lambda}{\mu}]_+\frac{\tilde{\mathbf{c}}_{(j)}}{||\tilde{\mathbf{c}}_{(j)}||_2}$, [22].

Given that we can solve the multi-block multi-class classification with group structured sparsity, we can also find the solution for adaptive lasso formulation. We can reformulate problem (4) as the initial estimation problem [18]. Define

$$\begin{aligned}&x_{new,l,i}=x_l||\tilde{\mathbf{w}}_{init,(\bmod (l,p))}||_2,\ l=1,..,rp,\ i=1,...,n\\ &\tilde{w}_{new,(j)}=\tfrac{\tilde{\mathbf{w}}_{(j)}}{||\tilde{\mathbf{w}}_{init,(j)}||_2},\ j=1,...,p\end{aligned}$$

then the adaptive lasso for sparse multi-block multi-class classification can be formulated as

$$\min_{F_{new}}\frac{1}{n}\sum_{i=1}^{n}V(F_{new},\mathbf{x}_{new,i})+\lambda_{adapt}R(F_{new}).$$

TABLE I
SIMULATION MODEL 1, FIVE-CLASS EXAMPLE, WITH $p = 1000$. CZ: NUMBER OF CORRECT ZEROS IN THE MULTI-CLASS CLASSIFIER, IZ: NUMBER OF INCORRECT ZEROS IN THE CLASSIFIER.

method	error rate	number of var. (CZ,IZ)
the ideal classifier	0	2 (998,0)
1. unified linear SVM	0.61	1000 (0,0)
2. sparse multi-class SVM	0.57	47 (952.5,0.5)
3. sparse multi-class SVM, prescreen	0.50	13.95 (985.8, 0.25)

The algorithm for sparse multi-block multi-class classification can be applied to this problem.

III. RESULTS

Fist, we implement a simulation model in [16], the five-class example. The model has independent variables with dimension p, and the first two variables are generated according to $N(\boldsymbol{\mu}_k, \sigma_1^2 I_2)$, where

$$\boldsymbol{\mu}_k = 2(\cos([2k-1]\pi/5), \sin([2k-1]\pi/5)),\ k = 1, 2, 3, 4, 5.$$

The remaining $p - 2$ variables are generated independently from $N(0, \sigma_2^2)$, and $\sigma_1 = \sqrt{2}$, $\sigma_2 = 1$. 250 samples are generated evenly from the model for training, another 250 samples for tunning the regularization parameter, and 50,000 samples for the test set. We compare the proposed sparse multi-class classification with the unified multi-class classifier in [10]. We also test the proposed algorithm with prescreening each pairwise binary classifications, i.e., the input variables to the multi-class classifier are the union of the variables selected by any pairwise binary classification. The entire experiment is repeated for 20 trials. We find that when $p < n$, the unified multi-class classifier without variable selection performs the best. However, when $p > n$, the multi-class classifier with structured variable selection outperforms the non-sparse classifiers. The results for $p = 1000$ are listed in Table I.

We the apply the sparse multi-class classification to classify gene expression of subjects over time. The data was collected from a challenge study where serial peripheral blood samples were acquired from a population of subjects inoculated with live (H3N2) flu virus [23]. The objective is to classify a sample (Affymetrix gene chip) into one of the three post-inoculation classes: the uninfected, the pre-infection and the acute-infection. The measurements right before inoculation are treated as the reference chips, whereas the ones after inoculation are the target chips . This is a two-block data, in which $p = 12023$, and $r = 2$. We prescreen the genes by pairwise classifications, and test the sparse multi-class classifiers, Figure 1. The majority classes are down sampled by limiting the number of samples per class per subject to be the same for all classes to overcome the imbalanced difficulty.

The performance of the proposed method is presented in Table II, compared with some classic classification methods. One subject is left out as the test set, and the rest as the training

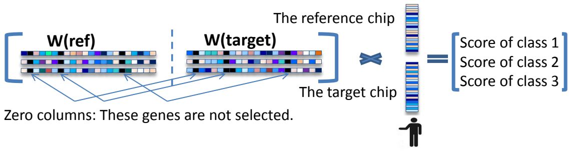

Fig. 1. Multi-block group structures for multi-class classification. The figure shows a multi-class classifier (K=3) matrix W, and the 2 block molecular data. The classifier matrix is divided into two blocks, denoted as W(ref) and W(target), which are associated with the reference sample and the target sample respectively.

TABLE II
Classification results by classic methods and the proposed sparse multi-class SVM.

methods	error rate	number of genes
linear SVM, one v.s. one	0.47	12023
unified linear SVM	0.38	12023
sparse multi-class SVM, prescreen	0.189	90.122

set. The parameters for all the methods are selected by 2-fold cross validation, and the samples are grouped by subjects, i.e., samples from the same subject should exist in the same set for cross validation. We take the z-scores on the training set, and standardize the variables in the test set accordingly. All the results are presented as averages over 4 repeated trials. Each trial evaluates the performance by leaving one subject out as the test set until all the subjects have been tested. Our results on the sparse multi-block multi-class problem show better performance than the classic approaches. This demonstrates the importance of the sparsity constraints.

IV. Conclusions

This paper develops a new framework for learning a classifier from a population of personalized serial samples. We derive a new variable splitting method for training a multi-class support vector machine with mixed L_1 and L_2 norm penalties that performs variable selection for optimal classification. Application of the classifier in the previous section shows significant improvement in the accuracy of classification of stages of host immune response of infected and uninfected subjects. The group sparsity penalty greatly reduces the number of variables and selects the most important ones for the classification task.

The method can be applied to other high-dimensional multi-block multi-class classification problems in bioinformatics and predictive health and disease tasks. By quantitative comparison of a person's current expression profile to that observed at previous times a more accurate health assessment can be made and more interpretable biomarkers can be discovered.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA), under the Predicting Health and Disease (PHD) program.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[3] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," *Advances in neural information processing systems*, vol. 16, no. 1, pp. 49–56, 2004.

[4] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: a stepwise procedure for building and training a neural network," in *Neurocomputing*. Springer, 1990, pp. 41–50.

[5] U. Kreßel, "Pairwise classification and support vector machines," in *Advances in kernel methods*. MIT Press, 1999, pp. 255–268.

[6] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[7] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proceedings of the seventh European symposium on artificial neural networks*, vol. 4, no. 6, 1999, pp. 219–224.

[8] E. Bredensteiner and K. Bennett, "Multicategory classification by support vector machines," *Computational Optimization and Applications*, vol. 12, no. 1, pp. 53–79, 1999.

[9] Y. Guermeur, "Combining discriminant models with new multi-class svms," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 168–179, 2002.

[10] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *J. Machine Learning Research*, vol. 2, pp. 265–292, 2002.

[11] Y. Liu and X. Shen, "Multicategory ψ-learning," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 500–509, 2006.

[12] L. Wang and X. Shen, "On l 1-norm multiclass support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 583–594, 2007.

[13] lldiko E. Frank and J. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[15] L. Wang and X. Shen, "On l 1-norm multiclass support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 583–594, 2007.

[16] H. Zhang, Y. Liu, Y. Wu, and J. Zhu, "Variable selection for the multicategory svm via adaptive sup-norm regularization," *Electronic Journal of Statistics*, vol. 2, pp. 149–167, 2008.

[17] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[18] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[19] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *Image Processing, IEEE Transactions on*, vol. 19, no. 9, pp. 2345–2356, 2010.

[20] S. Keerthi, S. Sundararajan, K. Chang, C. Hsieh, and C. Lin, "A sequential dual method for large scale multi-class linear svms," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 408–416.

[21] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[22] A. Puig, A. Wiesel, and A. Hero, "A multidimensional shrinkage-thresholding operator," in *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*. IEEE, 2009, pp. 113–116.

[23] Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Øien, M. T. McClain, J. B. Varkey, B. Nicholson, L. Carin, S. Kingsmore, C. W. Woods, G. S. Ginsburg, and A. O. Hero, III, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS genetics*, vol. 7, no. 8, p. e1002234, 2011.