# Regularized Least Squares Piecewise Multi-classification Machine

OLUTAYO O. OLADUNNI
BAO - Advanced Analytics and Optimization
IBM Global Business Services (GBS)
71 South Wacker Drive Chicago, IL 60606
USA
dr_o_olad@hotmail.com

*Abstract:* - This paper presents a Tikhonov regularization based piecewise classification model for multi-category discrimination of sets or objects. The proposed model includes a linear classification and nonlinear kernel classification model formulation. Advantages of the regularized multi-classification formulations include its ability to express a multi-class problem as a single and unconstrained optimization problem, its ability to derive explicit expressions for the classification weights of the classifiers as well as its computational tractability in providing the optimal classification weights for multi-categorical separation. Computational results are also provided to validate the functionality of the classification models using three data sets (GPA, IRIS, and WINE data).

*Key-Words:* - Piecewise, multi-class, multi-category discrimination, least squares, linear classifiers, nonlinear classifiers, linear system of equations

## 1 Introduction

The main idea underlying regularization theory is that an ill-posed problem or in particular, a problem of approximating the functional relation between $x$ and $y$ given a finite number of $l$ points $(x_i, y_i)_{i=1}^{l}$, can be formulated as a variational problem which contains both data and prior smoothness information. The smoothness of the function is taken into account by defining a smoothness functional $\varphi(f)$ in such a way that smaller values of the functional correspond to smoother functions. To find a function that is simultaneously close to the data and is also smooth, leads to the approximation problem arising from the minimization of the following quadratic functional [1], [2]:

$$H[f] = \frac{1}{l}\sum_{i=1}^{l}\left(y_i - f(x_i)\right)^2 + \lambda\varphi(f) \qquad (1)$$

for a fixed λ referred to as a regularization parameter. The first term is an $L_2$ loss function for empirical risk that enforces closeness of the data (reducing the empirical error or rate of misclassification). The second term called stabilizer enforces smoothness (generalization ability), and the regularization parameter controls the tradeoff between minimizing the generalization ability and the empirical error.

Support Vector Machines (SVMs), developed by Vapnik [3], [4] have been successfully applied to a wide range of problems. The SVM model as originally proposed requires the construction of several binary SVM classifiers to solve a multi-class problem. Although theoretically this scheme may seem practical, it becomes increasingly tedious to continue to perform the same repeated procedure of providing a solution for all independent classification models as this tends to increase the time used to obtain all solutions. The proposed least squares piecewise multi-classification model addresses this issue by solving the underlying multi-class problem as a linear system of equations that originates from a single optimization problem.

Most developed classification models are for discriminating between two classes. To address the problem of multi-classification, researchers have in the past adopted methods which involve solving $k$ SVM models (one-against-all (OAA) method) to produce $k$ classifiers [10] or solving $k(k-1)/2$ SVM models (one-against-one (OAO) method) to produce $k(k-1)/2$ classifiers; $k$ is the number of classes. Hsu & Lin [11] developed a decomposition strategy and made a comparison of the above methods. Methods include the OAA, OAO, and Direct Acyclic Graph SVM [12]. It was reported that the OAO method and DAGSVM are more suitable for practical use, and that for large scale problems, methods that consider all data at once, in general, use fewer support vectors.

In expressing and solving a multi-class problem as a single optimization problem the following models were suggested [13], [14], [15], [16], [17], [18], [19] and [20].

The objective of this study is aimed towards the reformulation of a piecewise multi-classification model using a least squares framework. This formulation leads to a linear system of equations which is smaller in size than the proposed least squares piecewise multi-classification model by Oladunni & Trafalis [18]. Benefits of this formulation include the expression of a multi-class problem as a single and unconstrained optimization problem, the explicit expressions for the classification weights of the classifiers, its computational tractability in providing the optimal classification weights for multi-categorical separation, its ability to provide solutions without the use of specialized solver-software, and a reduced linear system of equations.

The expression of the proposed method as a single optimization presents a less tiresome approach to attaining a solution to multi-class problems as opposed to solving several binary problems, which can be a time consuming effort. This approach offers a more succinct way to generating multi-class solutions. Furthermore, the proposed formulation leads to a strongly convex objective function that plays a key role in obtaining the optimal solution. In the proposed formulation, the fundamental change from other methods is the replacement of inequality constraints with equality. This modification, even if very simple, changes the nature of the optimization problem significantly. It turns out that one can write exact expressions/solutions to the problem in terms of the problem data as shown in the subsequent sections, whereas it is impossible to do that in the previous linear and quadratic programming formulations because of their combinatorial nature that require optimization solvers.

This paper is organized as follows. In section 2, a short description of a piecewise multi-classification problem is presented. In section 3, a description of the proposed regularized least squares piecewise multi-classification machine (RLSM) model is given. In section 4, several data sets for numerical testing are described. In section 5, computational results are presented, and section 6 concludes the paper.

## 2 Piecewise Multi-classification Formulation

Consider $m$ vectors in $R^n$, belonging to $k$ classes ($k$>2), where each class comprises of $m_i$ vectors such that $\sum_{i=1}^{k} m_i = m$. Assume that the set of vectors belonging to the $k$ classes are piecewise-linearly separable, i.e., there exist $w^i \in R^n$ and $\gamma^i \in R$ such that

$$A^i w^j - \gamma^i e^i > A^i w^j - \gamma^j e^i,\ i, j = 1,...,k,\ i \neq j, \quad (2)$$

where $A^i$ is an $m_i \times n$ matrix whose rows are the input data points in the $i^{th}$ class, $e^i$ is a vector of ones with $m_i$ elements and the difference between $w^i$ and $w^j$ is the normal vector perpendicular to the optimal hyperplane. The location of the optimal hyperplane relative to the origin are determined by the difference value of $\gamma^i$ and $\gamma^j$. In canonical form

$$x^T(w^i - w^j) - e^i(\gamma^i - \gamma^j) \geq 1,\ i, j = 1,...,k,\ i \neq j. \quad (3)$$

The bounding plane separating classes $i$ and $j$ is (see Fig. 1)

$$x^T(w^i - w^j) = (\gamma^i - \gamma^j). \quad (4)$$

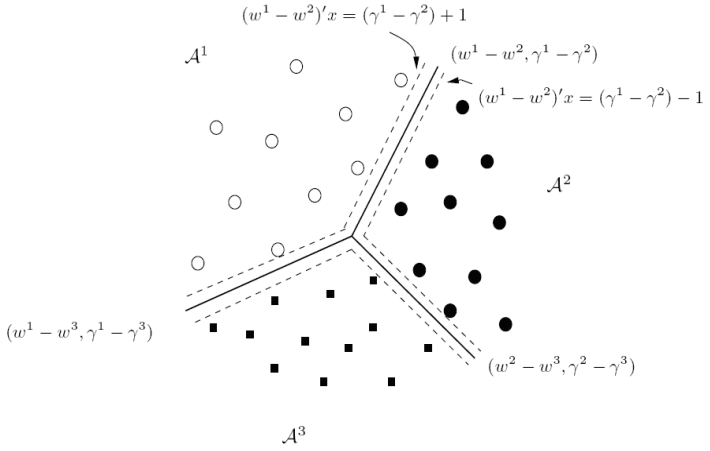


Fig 1: Piecewise-linear separator with margins for three classes [13]

Adding a regularization term $\frac{1}{2}\sum_{i=1}^{k}\left\|w^i\right\|^2$ to the objective function leads to the formulation of a piecewise linearly separable multi-classification problem (see Fig. 1):

$$\min_{w,\gamma} \ \frac{1}{2}\sum_{i=1}^{k}\left\|w^i\right\|_2^2$$
$$s.t.\ A^i(w^i - w^j) - e^i(\gamma^i - \gamma^j) - e \geq 0 \qquad (5)$$
$$i, j = 1,\cdots,k \quad i \neq j$$

To classify a new point $x$, compute $g_i(x)$

$$g_i(x) = x^T w^i - \gamma^i, \qquad (6)$$

and find $i$ such that $g_i(x)$ is maximized, i.e., $g(x) = \max_{i=1,..,k} g_i(x)$, where $g(x)$ is a decision function.

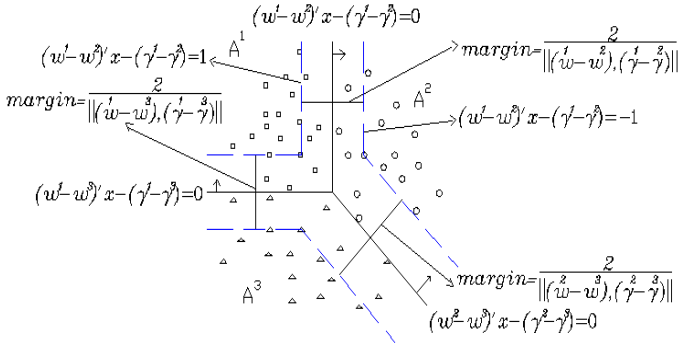


Fig. 2: Least squares piecewise-linear separator with margins for three classes

## 3 Regularized Least Squares Multi-classification Machine: Piecewise Formulation

In this section, a regularized least squares multi-category machine (RLSM) for linear and nonlinear classification is presented. The derivation of the RLSM is based on problem (5), where the norm is minimized simultaneously with the sum of square error and the relative location of the origin $\gamma$. The objective is to make a very simple, but very fundamental change in the formulation (5), specifically replace the inequality constraint by an equality constraint (see equation 7).

This modification makes it possible to reduce the optimization problem to a linear system of equations, as well as derive explicit expressions for the solution parameters. Another interesting observation is the increase in margin, whereas it is a narrower margin for formulation (5). This happens because the hyperplane $x^T\left(w^i - w^j\right) = \left(\gamma^i - \gamma^j\right) + 1$ associated with class $i$ and $x^T\left(w^i - w^j\right) = \left(\gamma^i - \gamma^j\right) - 1$ associated class $j$ move closer to their respective data point cluster (see Fig. 2).

### 3.1 Linear Least Squares Piecewise Multi-classification Machine

Consider a problem of classifying data sets in $R^n$ that are represented by a data matrix $A^i \in R^{m_i \times n}$, where $i = 1,..,k$ ($k \geq 3$ classes). $A^i$ is an $m_i \times n$ matrix whose rows are the input data points in the $i^{th}$ class. This problem can be modeled through the following optimization problem (see Fig. 2):

$$\min_{w,\gamma,\xi} \ \frac{\lambda}{2}\left(\sum_{i=1}^{k}\left\|w^i\right\|_2^2 + \sum_{i=1}^{k}\left(\gamma^i\right)^2\right) + \frac{1}{2}\sum_{i=1}^{k}\sum_{t=1}^{m_i}\left(\xi_t^{ij}\right)^2$$
$$s.t.\ A_t^i(w^i - w^j) - (\gamma^i - \gamma^j) = 1 - \xi_t^{ij} \qquad , \qquad (7)$$
$$i, j = 1,\cdots,k \quad i \neq j,\ t = 1,\cdots,m_i$$

where $\lambda$ is the regularization parameter between minimizing the empirical error of the training set (data) and maximizing the margin (minimizing generalization ability), $t$ denotes the $t^{th}$ row of data matrix $A^i$, and $\xi_t^{ij}$ is an error slack variable accounting for the empirical error of the training set. Here is a 3 classes problem ($k$ = 3) rewritten in matrix notation.

$$\bar{A} = \begin{pmatrix} A^1 & -A^1 & 0 \\ A^1 & 0 & -A^1 \\ -A^2 & A^2 & 0 \\ 0 & A^2 & -A^2 \\ -A^3 & 0 & A^3 \\ 0 & -A^3 & A^3 \end{pmatrix} \quad \bar{E} = \begin{pmatrix} -e^1 & e^1 & 0 \\ -e^1 & 0 & e^1 \\ e^2 & -e^2 & 0 \\ 0 & -e^2 & e^2 \\ e^3 & 0 & -e^3 \\ 0 & e^3 & -e^3 \end{pmatrix}, \qquad (8)$$

where $A^i \in R^{m_i \times n}$ is matrix whose rows are input data belonging to a class, and $e^i \in R^{m_i \times 1}$, $i = 1,..,k$ ($k = 3$ classes) is a vector of ones. So when k > 2, we simply adjust matrices $\bar{A}$ and $\bar{E}$. Here is a $k$ class problem rewritten in matrix notation.

Let

$$\bar{A}=\begin{pmatrix} A^1 & -A^1 & 0 & 0 & \cdots & 0 \\ A^1 & 0 & -A^1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & 0 & \ddots & \vdots \\ A^1 & 0 & \cdots & \cdots & 0 & -A^1 \\ -A^2 & A^2 & 0 & 0 & \cdots & 0 \\ 0 & A^2 & -A^2 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & A^2 & 0 & \cdots & \cdots & -A^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -A^k & 0 & \cdots & \cdots & 0 & A^k \\ 0 & -A^k & 0 & \cdots & 0 & A^k \\ \vdots & 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & -A^k & A^k \end{pmatrix} \quad (9)$$

where $A^i \in R^{m_i \times n}$. The matrix $\bar{A}$ has $(k-1)\sum_{i=1}^{k} m_i$ rows and *kn* columns.

$$\bar{E}=\begin{pmatrix} -e^1 & e^1 & 0 & 0 & \cdots & 0 \\ -e^1 & 0 & e^1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & 0 & \ddots & \vdots \\ -e^1 & 0 & \cdots & \cdots & 0 & e^1 \\ e^2 & -e^2 & 0 & 0 & \cdots & 0 \\ 0 & -e^2 & e^2 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & -e^2 & 0 & \cdots & \cdots & e^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e^k & 0 & \cdots & \cdots & 0 & -e^k \\ 0 & e^k & 0 & \cdots & 0 & -e^k \\ \vdots & 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & e^k & -e^k \end{pmatrix}, \quad (10)$$

where $e^i \in R^{m_i \times 1}$ is a vector of ones. The matrix $\bar{E}$ has $(k-1)\sum_{i=1}^{k} m_i$ rows and *k* columns.

Below is the matrix expression for problem (7)

$$\min_{w,\gamma,\xi} \ \frac{\lambda}{2}\left\|\begin{pmatrix} w \\ \gamma \end{pmatrix}\right\|_2^2 + \frac{1}{2}\|\xi\|_2^2 \ , \quad (11)$$

$$s.t. \ \ \bar{A}w + \bar{E}\gamma - e + \xi = 0$$

where $w=\left[\left(w^1\right)^T,\left(w^2\right)^T,....,\left(w^k\right)^T\right]^T$,

$\gamma=\left[\gamma^1,\gamma^2,....,\gamma^k\right]^T$, $e$ is a vector of ones with appropriate dimension, and

$\xi=\left[\left(\xi^{12}\right)^T,\left(\xi^{13}\right)^T,....,\left(\xi^{1k}\right)^T,\left(\xi^{21}\right)^T,\left(\xi^{23}\right)^T,....,\left(\xi^{k(k-1)}\right)^T\right]^T$.

Problem (11) is a constrained optimization problem. Using the concept of penalty functions [21], problem (11) can be rewritten as an unconstrained optimization problem given below:

$$\min_{w,\gamma} f(w,\gamma) = \frac{\lambda}{2}\left\|\begin{pmatrix} w \\ \gamma \end{pmatrix}\right\|_2^2 + \frac{1}{2}\left\|\bar{A}w + \bar{E}\gamma - e\right\|_2^2 . \quad (12)$$

Setting $v=\begin{bmatrix} w^T & \gamma^T \end{bmatrix}^T$ and defining *H* as:

$$H=\begin{bmatrix} \bar{A} & \bar{E} \end{bmatrix}, \quad (13)$$

problem (12) can be written succinctly as:

$$\min_{v} f(v) = \frac{\lambda}{2}\|v\|_2^2 + \frac{1}{2}\|Hv - e\|_2^2 . \quad (14)$$

Problem (14) is an unconstrained optimization problem for piecewise linear multi-classification. The tradeoff constant $\lambda$ is run through a series or range of values to achieve the best result. If its value increases then the minimum norm is achieved, but at the expense of having a higher training residual error (empirical error). The first term of problem (14) ensures the smoothness of the function, i.e., similar inputs have similar outputs; the second term ensures that the data points in each class are as close as possible to each other.

Problem (14) is a convex unconstrained optimization problem which has a minimum point. The minimizing point of problem (14) is the solution to the following optimality condition(s) of $f(v)$ set to equal zero:

$$\frac{df}{dv} = \lambda v + H^T H v - H^T e = 0 . \quad (15)$$

From equation (15) the following expression is obtained for $v$:

$$v = \left(\lambda I + H^T H\right)^{-1} H^T e . \quad (16)$$

Setting

$$M = \lambda I + H^T H\ ,\ t = H^T e, \tag{17}$$

the solution to problem (14) can be written succinctly as:

$$v = M^{-1} t\ , \tag{18}$$

and its corresponding linear system of equations is:

$$Mv = t,\ \text{where}\ v = \begin{bmatrix} w^T & \gamma^T \end{bmatrix}^T . \tag{19}$$

To classify a new point *x*, compute

$$g_i(x) = x^T w^j - \gamma^j , \tag{20}$$

and find $i$ such that $g_i(x)$ is maximized, i.e., $g(x) = \max_{i=1,..,k} g_i(x)$, where $g(x)$ is a decision function.

Assuming the matrix *M* in Equation (18) is invertible then it provides a solution to the linear system of equations in (19). Methods for solving the linear system of equations include matrix decomposition methods and/or iterative based methods [21], [22]. Its solution involves the evaluation of a smaller dimensional matrix (*M*) of magnitude $(kn+k)\times(kn+k)$. Below is procedure for the linear RLSM.

ALGORITHM 3.1.1 - **Linear RLSM**: Given a data set in $R^n$ that is represented by a matrix $A^i \in R^{m_i \times n}$, where $i = 1,..,k$ classes. Classification weights *w* and $\gamma$ for linear classifiers are computed as follows.

Step 1: Define *H* by (13).
Step 2: Compute *M* and *t* from (17).
Step 3: Determine *v* from (18).
Step 4: Classify a new point *x* by (20).

Matrix methods or iterative methods are employed to obtain the solution in step 3.

## 3.2 Nonlinear RLSM

To formulate the nonlinear counterpart of the piecewise linear classification model, the primal variable *w* is replaced by its equivalent dual representation as shown below:

$$w = \bar{A}^T \alpha\ , \tag{21}$$

where $\alpha$ is the vector of dual variables. We have the following piecewise multi-classification Tikhonov regularization formulation for linear separation in dual space by substituting *w* in (21) into (12).

$$\min_{\alpha,\gamma} f(\alpha,\gamma) = \frac{\lambda}{2}\left\| \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \right\|_2^2 + \frac{1}{2}\left\| \bar{A}\bar{A}^T \alpha + \bar{E}\gamma - e \right\|_2^2 , \tag{22}$$

where

$$\alpha^T = \left[ \left(\alpha^{12}\right)^T, \left(\alpha^{13}\right)^T, ...., \left(\alpha^{1k}\right)^T, \left(\alpha^{21}\right)^T, \left(\alpha^{23}\right)^T, ...., \left(\alpha^{k(k-1)}\right)^T \right]$$

and $\gamma = \left[ \gamma^1, ..., \gamma^k \right]^T$.

However, to obtain nonlinear classifiers it is essential to carry out a nonlinear mapping from the input space to a feature space using a mapping function $\phi : R^n \to F$ [23], [24]. Since we do not know the mapping function $\phi$, a kernel function is employed to implicitly compute the inner products of the input vectors in feature space. Depending on the specific kernel function chosen, the resulting kernel matrix defines the similarity or dissimilarity of the input vectors. The kernel function can define a distance in the input space. Two of the most widely used kernel functions are as follows [24]:

$k(x_i, x) = (x_i^T x + 1)^P$, where *p* is the degree of polynomial for the polynomial (poly) kernel function.

$k(x_i, x) = \exp(-\sigma \|x_i - x\|^2)$, σ is the width or spread for the radial basis function (rbf) kernel function.

Given $A \in R^{m\times n}$ and $B \in R^{n\times k}$, the ***kernel*** $K(A,B)$ *maps* $R^{m\times n} \times R^{n\times k}$ *into* $R^{m\times k}$. The kernel function satisfies Mercer's condition and therefore the kernel matrix is a symmetric positive semi-definite (PSD) matrix [23], [24].

Problem (22) is a piecewise linear classification problem, because $\bar{A}\bar{A}^T$ can be considered as a linear kernel. In order to generalize to piecewise nonlinear classifiers, the linear kernel is replaced by a general nonlinear kernel $K(\bar{A}, \bar{A}^T)$ and problem (22) becomes:

$$\min_{\alpha,\gamma} f(\alpha,\gamma) = \frac{\lambda}{2}\left\| \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \right\|_2^2 + \frac{1}{2}\left\| K(\bar{A}, \bar{A}^T)\alpha + \bar{E}\gamma - e \right\|_2^2 . \tag{23}$$

Setting $c = \begin{bmatrix} \alpha^T & \gamma^T \end{bmatrix}^T$, $K = K(\bar{A}, \bar{A}^T)$ and defining *G* as:

$$G = \begin{bmatrix} K & \bar{E} \end{bmatrix}, \tag{24}$$

problem (23) can be written succinctly as:

$$\min_{c} f(c) = \frac{\lambda}{2}\|c\|_2^2 + \frac{1}{2}\|Gc - e\|_2^2 . \tag{25}$$

Problem (25) is a convex unconstrained optimization problem for piecewise nonlinear multi-classification. The minimizing point of problem (25) is the solution to the following optimality condition(s) of $f(c)$ set to equal zero:

$$\frac{df}{dc} = \lambda c + G^T G c - G^T e = 0 . \quad (26)$$

From equation (26) the following expression is obtained for $c$:

$$c = \left(\lambda I + G^T G\right)^{-1} G^T e . \quad (27)$$

Setting

$$Q = \lambda I + G^T G , \; z = G^T e , \quad (28)$$

the solution to problem (22) can be written succinctly as:

$$c = Q^{-1} z . \quad (29)$$

and its corresponding linear system of equations is:

$$Qc = z, \text{ where } c = \left[\alpha^T \quad \gamma^T\right]^T . \quad (30)$$

Solving for $w^i$ in summation notation we get:

$$w^i = \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{t=1}^{m_i} \alpha_t^{ij} (A_t^i)^T - \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{t=1}^{m_j} \alpha_t^{ji} (A_t^j)^T . \quad (31)$$

Therefore to classify a new point *x*, compute

$$g_i(x) = \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{t=1}^{m_i} \alpha_t^{ij} K\left(x^T, (A_t^i)^T\right) - \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{t=1}^{m_j} \alpha_t^{ji} K\left(x^T, (A_t^j)^T\right) - \gamma^i , \quad (32)$$

and find $i$ such that $g_i(x)$ is maximized, i.e., $g(x) = \max_{i=1,..,k} g_i(x)$, where $g(x)$ is a decision function. Assuming the matrix $Q$ in Equation (29) is invertible then it provides a solution to the linear system of equations in (30). Below is procedure for the nonlinear RLSM.

ALGORITHM 3.2.1 - **Nonlinear RLSM**: Given a data set in $R^n$ that is represented by a matrix $A^i \in R^{m_i \times n}$, where $i = 1,..,k$ classes. Classification weights $\alpha$ and $\gamma$ for nonlinear classifiers are computed as follows.

Step 1: Choose a kernel function $K = K(\bar{A}, \bar{A}^T)$.
Step 2: Define *G* by (24).
Step 3: Compute *Q* and *z* from (28).
Step 4: Determine *c* from (29).
Step 5: Classify a new point *x* by (32).

Matrix methods or iterative methods are employed to obtain the solution in step 4.

# 4 Numerical Testing

In this section, the description of three data sets trained on a regularized least squares multi-category machine (RLSM) for discriminating between *k* classes is presented. Below are the three data sets of interest:

**Table 1. List of Datasets.**

| Dataset | No. of Attributes | No. of Points | No. of Classes |
|---|---|---|---|
| GPA | 2 | 85 | 3 |
| IRIS | 4 | 150 | 3 |
| WINE | 13 | 178 | 3 |

Admission Data for Graduate School of Business: The Admission data set [17], [18], [32] uses the undergraduate grade point average (GPA) and graduate management aptitude test (GMAT) scores to help determine which applicants should be admitted to the school's graduate program. There are 85 instances (points) and 2 attributes (features), 43 data points used as training samples and 42 data points used as test samples. The distribution of instances with respect to their class is as follows: 28 instances in class 1 (not admitted), 26 instances in class 2 (borderline), and 31 instances in class 3 (admitted).

Iris flower Data: The Iris data set [32] uses the sepal length, sepal width, petal length, and petal width to help discriminate between three species of iris flower. There are 150 instances (points) and 4 attributes (features), 75 data points used as training samples and 75 data points used as test samples. The distribution of instances with respect to their class is as follows: 50 instances for each of the three classes. Class 1 belongs to the iris setosa specie, class 2 is the iris versicolor specie, and class 3 is the iris virginica specie.

Wine Recognition Data: The Wine data set uses the chemical analysis of wine grown in the same region in Italy to help discriminate between three different cultivars. There are 178 instances (points) and 13 attributes (features), 90 data points used as training samples, 88 data points used as test samples. The distribution of instances with respect to their class is as follows: 59 instances belong to class 1, 71

instances belong to class 2 and 48 instances belong to class 3. The WINE data set were obtained from the UCI Repository of Machine Learning Databases and Domain Theories [33].

## 5 Computational Results

In this section, the result of the analyzed data set is presented and discussed. The LRLSM and NRLSM models were used to train the data sets. A random sample validation was employed to validate the proposed models. In the random sample validation, fifty percent of the data set was drawn randomly and trained on both proposed models, while the other 50% was used as test samples. Ten random samples of each data set were trained and tested, and the misclassification error for each test sample was recorded. The analysis of the NRLSM model was performed using a polynomial and RBF kernel functions.

All classification analysis were implemented using MATLAB [34], and comparisons were made by evaluating a performance parameter (misclassification error) defined below:

$$error = 1 - \left( \frac{Total\ number\ of\ correctly\ classified\ points}{Total\ number\ of\ observed\ points} \right), \quad (33)$$

where error represents the overall misclassification error rate, i.e., the fraction of misclassified points for a test sample of a given data set. For 100% classification or correctness, $error = 0$. Results of the data sets trained on the LRLSM (14) and NRLSM (25) models are shown in Tables 2 – 4.

**Table 2. Test error rates of LRLSM and NRLSM on GPA, IRIS and WINE data.**

| Data Set | LRLSM | NRLSM-RBF | NRLSM-Poly |
|---|---|---|---|
| GPA | 0.0476 | **0.0238** | **0.0238** |
| IRIS | **0.0533** | 0.0933 | 0.2 |
| WINE | **0.0000** | 0.1818 | 0.0114 |

Table 2 presents the average test error rate results for the LRLSM & NRLSM models respectively, based on ten runs (random sample validation) for the GPA, IRIS, WINE data. The linear model (LRLSM) reports a better error rate than the nonlinear model (NLRLSM) in 2 out 3 three data sets. For the GPA data set, the NRLSM model reports the best error rate of 0.0238. It appears that either the polynomial or RBF kernel function will suffice in the classification of the GPA data set. The best error rate for the IRIS data set is 0.0533, which was attained by the LRLSM model. The best error rate for the WINE data set is 0, which was also attained by the LRLSM model.

The results of the regularized piecewise multi-class models were further compared with other results from other multi-class learning algorithms in the literature [13], [15] and [19]. The WINE data set results were compared to the piecewise multi-class learning algorithms from Bredensteiner & Bennett [13], which include the multi-class robust linear programming & *k*-class robust linear programming (M-RLP & *k*-class), and multi-class support vector machines & *k*-class support vector machines (M-SVM & *k*-SVM). All three data set results were compared to regularized pair-wise models of Suykens & Vandewalle [15] least squares multi-class support vector machines and the multi-class learning algorithms from Oladunni & Trafalis [19], which include three Tikhonov regularization ($T_R$) classification methods, linear multi-class $T_R$SVM ($L_MT_R$SVM), nonlinear multi-class $T_R$ kernel machine ($NL_MT_R$KM), and reduced kernel multi-class $T_R$ machine ($RK_MT_R$M) models.

In Table 3 are the error rates reported in the work of Oladunni & Trafalis [19] (regularized pair-wise models) in comparison with the regularized least squares piecewise multi-class models. In bold are the lowest error rates.

The regularized least squares piecewise multi-class models (LRLSM and NRLSM) reports a better error rate than the regularized pair-wise models in 2 out 3 three data sets. Comparing the pair-wise multi-class algorithms, the nonlinear models outperform the linear models. For the GPA data set, the best error rate (0.0637) was obtained by the $RK_MT_R$M model and it is considerably worse than the piecewise models. For the IRIS data set, the best error rate (0.0178) was obtained using a $NL_MT_R$KM model and it is considerably better than the piecewise models. For the WINE data set, the best error rate (0.0152) was obtained using the $NL_MT_R$KM and it is considerably worse than the piecewise models, especially in the case of the LRLSM, which has a zero test error rate.

**Table 3. Comparison of test error rates for piecewise and pair-wise methods.**

| Method | GPA | IRIS | WINE |
|---|---|---|---|
| LRLSM | 0.0476 | 0.0533 | **0.0000** |
| NRLSM-RBF | **0.0238** | 0.0933 | 0.1818 |
| NRLSM-Poly | **0.0238** | 0.2000 | 0.0114 |
| $L_MT_RSVM$ | 0.0714 | 0.0222 | 0.0227 |
| LS-MSVM | 0.0714 | 0.0222 | 0.0303 |
| $NL_MT_RKM$ | 0.0714 | **0.0178** | 0.0152 |
| $RK_MT_RM$ | 0.0635 | 0.0222 | 0.0227 |

**Table 4. Comparison of WINE data set test error rates for piecewise methods.**

| Method | Error Rate |
|---|---|
| LRLSM | **0.0000** |
| NRLSM-RBF | 0.1818 |
| NRLSM-Poly | 0.0114 |
| M-RLP | 0.0899 |
| k-RLP | 0.0056 |
| M-SVM | 0.0281 |
| k-SVM | 0.0056 |
| LPPMSVM | 0.0114 |

In Table 4 are the error rates reported in the work of Bredensteiner & Bennet [13] (M-RLP, *k*-RLP, M-SVM and *k*-SVM models) and Oladunni & Singhal [20] (LPPMSVM model) in comparison with the regularized least squares piecewise multi-class models. In bold are the lowest error rate(s).
Although the *k*-class robust linear programming (*k*-RLP) and *k*-class support vector machines (*k*-SVM); both with error rates of 0.0056, come close to the results obtained with the LRLSM model, analysis shows that the LRLSM model outperforms all the other models[1]. This demonstrates the potential of the regularized least squares piecewise multi-class models, and above all shows the potential of the linear piecewise classification functions over the nonlinear piecewise classification functions.

[1] This conclusion assumes similarities between the experimental setup used for analysis here and that used in the Bredensteiner & Bennet [13] paper.

## 6 Conclusions

In this paper, we have proposed two easy to implement regularized least squares multi-class models for linear and nonlinear piecewise multi-classification. This was done by reformulating the MSVM using equality constraints and concept of penalty functions, which leads to the approximation problem arising from the minimization of a quadratic functional. These piecewise multi-class formulations are single unconstrained optimization problems for which solutions can be obtained via solving a linear system of equations.

A computational study of the regularized least squares piecewise multi-class models were performed on three data sets. In general we found that all multi-class methods generalized. The nonlinear regularized least squares piecewise multi-class model (NRLSM) performed best on the GPA data set. The $NL_MT_RKM$ model which is a regularized pair-wise formulation performed best on the IRIS data set, and the linear regularized least squares piecewise multi-class model (LRLSM) performed best on the WINE data set, but the *k*-RLP and *k*-SVM models also performed well.

In order to make computations of the nonlinear regularized least squares piecewise multi-class model more tractable, future work should include the investigation of a reduced kernel formulation for nonlinear classifiers for large scale problems.

*References:*

[1] A.N. Tikhonov and V.Y. Arsenin, *Solution of Ill-Posed Problems*. Washington D.C.: Winston, 1977.

[2] V.V. Ivanov, *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*. Nordhoff, International, 1976.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*. New-York: Springer-Verlag, 1995.

[4] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

[5] G. Fung and O.L. Mangasarian, "Multicategory proximal support vector machine classifiers," *Machine Learning*, vol. 59, pp. 77-97, 2005.

[6] G. Fung, O.L. Mangasarian, and J.W. Shavlik, "Knowledge-based support vector machine classifiers," in *Proceedings NIPS* 2002, Vancouver, BC, December 10-12, 2002.

[7] G. Fung, O.L. Mangasarian, and J.W. Shavlik, "Knowledge-based nonlinear kernel classifiers,"

in Manfred Warmuth and Bernhard Scholkopf (eds), in *Proceedings Conference on Learning Theory (COLT/Kernel 03) and Workshop on Kernel Machines*, 2003, pp. 102-113, Washington, D.C., August 24 - 27.

[8] O.L. Mangasarian, J.W. Shavlik, and E.W. Wild, "Knowledge-based kernel approximation," *Journal of Machine Learning Research*, vol. 5, pp. 1127-1141, 2004.

[9] J.A.K. Suykens, L. Lukas, and J. Vandewalle, "Sparse approximation using least squares support vector machines," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'00)*, Geneva, Switzerland, pp. II 757-II 760, May, 2000.

[10] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *Journal of Machine Learning Research*, vol. 5, pp. 101-141, 2004.

[11] C-W. Hsu and C-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, pp. 415 – 425, 2002.

[12] J.C. Platt, N. Cristianini, and J. Shawe-Taylor, "Largin margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems*, vol. 12, pp. 547-553, MIT Press, 2000.

[13] E.J. Bredensteiner and K.P. Bennett, "Multicategory classification by support vector machines," *Computational Optimization and Applications*, vol. 12, pp. 53 – 79, 1999.

[14] J. Weston and C. Watkins, "Multi-class support vector machines," in M. Verleysen, editor, *Proceedings of ESANN99*, Brussels, D. Facto Press, 1999.

[15] J.A.K. Suykens and J. Vandewalle, "Multiclass least squares support vector machine classifers," in *Proceedings. of Joint Conf. on Neural Networks (IJCNN'99)*, Washington, D.C., 1999c.

[16] S. Szedmak, J. Shawe-Taylor, C.J. Saunders, and D.R. Hardoon, "Multiclass classification by $L_1$ norm support vector machine," in *Pattern Recognition and Machine Learning in Computer Vision Workshop*, Grenoble, France, May 02-04, 2004.

[17] T.B. Trafalis and O. Oladunni, "Pairwise multi-classification support vector machine: quadratic programming (QP-$P_A$MSVM) formulations," *WSEAS Transactions on Systems*, vol. 4, pp. 349-354, April 2005.

[18] O. Oladunni and T.B. Trafalis, "Least square multi-classification support vector machines: pairwise ($P_A$LS-MSVM) & piecewise ($P_I$LS-MSVM) formulations," *WSEAS Transactions on Circuits and Systems*, vol. 4, pp. 363-368, April, 2005.

[19] O.O. Oladunni and T.B. Trafalis, "A pairwise reduced kernel-based multi-classification Tikhonov regularization machine," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'06)*, IEEE Press, pp. 130 – 137, Vancouver, BC, Canada, July 2006, on CD-ROM.

[20] O.O. Oladunni and G. Singhal, "Piecewise Multi-Classification Support Vector Machines," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'09)*, IEEE Press, pp. 2323 – 2330, Atlanta, GA, June 2009, on CD-ROM.

[21] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty, *Nonlinear Programming – Theory and Algorithms*. John Wiley & Sons, Inc., 1993.

[22] J. M. Lewis, S. Lakshmivarahan, and S. Dhall, *Dynamic Data Assimilation*. Cambridge University Press, 2006.

[23] C.J.C. Burges, "A tutorial on support vector machines for pattern classification," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[24] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[25] T.B. Trafalis and H. Ince, "Support vector machine for regression and applications to financial forecasting," *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Como, Italy, July: 24-27, 2000.

[26] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349-358, 2001.

[27] B. Santosa, T. Conway, and T.B. Trafalis, "Knowledge based-clustering and application of multi-class SVM for genes expression analysis," *Intelligent Engineering Systems through Artificial Neural Networks*, vol. 12, pp. 391 – 395, 2002.

[28] Y-J. Lee, O.L. Mangasarian, and W.H. Wolberg, "Survival-time classification of breast cancer

patients," *Computational Optimization and Applications*, vol. 25, pp. 151-166, 2003.

[29] A.M. Malyscheff and T.B. Trafalis, "Support vector machines and the electoral college," in *Proceedings of the International Joint Conference on Neural Networks*, IEEE Press, pp. 2345-2348, Portland, Oregon, USA, 2003.

[30] T.B. Trafalis and O. Oladunni, "Single phase fluid flow classification via neural networks & support vector machine," *Intelligent Engineering Systems Through Artificial Neural Networks*, (C.H. Dagli, A.L. Buczak, D. L. Enke, M.J. Embrechts, and O. Ersoy, eds.), ASME Press, vol. 14, pp. 427-432, 2004.

[31] T.B. Trafalis, O. Oladunni, and D.V. Papavassiliou, "Two-phase flow regime identification with a multi-classification SVM model," *Industrial & Engineering Chemistry Research*, vol. 44, pp. 4414 – 4426, 2005.

[32] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistics Analysis*. New Jersey: Prentice Hall, 2002.

[33] P.M. Murphy and D.W. Aha, *UCI repository of machine learning databases*. [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Department of Information and Computer Science, University of California, Irvine, California, 1994.

[34] *MATLAB User's Guide*. The Math-Works, Inc., Natwick, MA 01760, 1994-2003. http://www.mathworks.com.