RODLSR: ROBUST DISCRIMINATIVE LEAST SQUARES REGRESSION MODEL FOR MULTI-CATEGORY CLASSIFICATION

Lingfeng Wang[1,2], *Shuaizheng Liu*[2] *and Chunhong Pan*[2]

1. Hunan Provincial Key Laboratory of Network Investigational Technology
2. NLPR, Institute of Automation, Chinese Academy of Sciences
{lfwang@nlpr.ia.ac.cn, susanna5lsz@163.com, and chpan@nlpr.ia.ac.cn}

ABSTRACT

Discriminative least squares regression (DLSR) is a simple yet effective method for multi-class classification. One problem of DLSR is that it is lack of robustness to outliers. In order to tackle this difficulty, in this paper, we propose a novel Robust DLSR (RoDLSR) model. The core idea behind RoDLSR is to find and further ignore the outliers among the support vector set. Specifically, we modify the regression targets of outliers by adding an additional item. As a result, the range of regression residuals can be controlled within predefined threshold. Extensive experiments evaluate the effectiveness of RoDLSR, especially on the corrupted databases.

***Index Terms*—** LSR, DLSR, Robust, Robust DLSR (RoDLSR), Support Vector

1. INTRODUCTION

Least squares regression (LSR) is widely adapted to statistical analysis. It matches the best function for the data by minimizing the square of error. Given a data set $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^d$ and a target set $\{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^c$, LSR can be define as

$$\min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{n} \|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\|_2^2 + \beta\|\mathbf{W}\|_F^2 \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ and $\mathbf{b} \in \mathbb{R}^c$ are to be estimated and the parameter β is a regularization parameter.

As above compact form and efficient achievement, LSR institutes its fundamental status in data processing and classification. And there are many derivatives of LSR which can be divided into two categories. One is revising the loss function, such as logistic regression (LR) [1], LASSO [2], ridge regression [3], SVM with the hinge loss [4], SVM with the square hinge loss [5], least squares(LS)SVM [6], [7], [8]. Above mentioned methods are binary classification models, so we adopt one-against-one, one-against-rest or ECOC [9] strategy and combine the independent obtained models for the multi-category classification. The other category is transforming the regression targets. The typical instances are recently proposed discriminative LSR (DLSR) [10], retargeted LSR (ReLSR) [11] and margin scalable DLSR (MSDLSR) [12] where ReLSR and MSDLSR are both based on DLSR. DLSR introduces a technique called ε-dragging to dislodge the regression targets towards opposite direction, for extending the interval between different classes. What's more, DLSR can tackle multi-category classification under a compact model. And experiment results testify DLSR performance better than traditional regression methods and SVM-based methods especially under the presence of sufficient training data.

However, DLSR is lack of robustness to the outliers and noises, which are common in realistic situation due to objective and subjective factors. The outliers can produce large residuals, which exert critical effects on the estimate of regression parameters. The ultimate regression results will stray from the correct one. The studies for robust regression sustain to develop well. The classical method is M-estimation [13] introduced by Huber, whose core idea is selecting suitable weight function to attach corresponding weight coefficient to every residual and minimizing the weighted residual sum of squares. Least trimmed squares(LTS) [14] and S-estimation [15] are other viable robust regression method. Besides above measures, we can replace the normal distribution with a heavy-tailed distribution,such as Bayesian robust regression [16]. However, when above mentioned robust approaches are applied to DLSR. we will obtain multiple models that are independent of each other. The computation process is complicated and the effects cannot satisfy our expectations, which demonstrates the above solutions don't suit for it.

In this paper, we propose a novel model called RobustDLSR (RoDLSR) to accomplish the robust regression. According to the exploration and comparison to the loss function of LSR and DLSR, we expect to formulate a unified and compact model to control the range of residuals. This paper utilizes positive slack variable ε_i in DLSR to find the outliers, and attaches an additional item to the regression targets of outliers, which assures residuals are below a predefined value and weaken the effects from the outliers. The propose approach is validated extensively on several datasets.

The outline of this paper is organized as follows. Section 2 gives a brief about DLSR and its variations. Section 3 mainly introduces the robust model proposed in this paper and its optimization solution. Section 4 exhibits the comparable and satisfying experimental results. Section 5 summarizes the core idea and discusses the further improvement of our method to conclude this paper.

2. DLSR AND ITS VARIATIONS

Some previous work impels us to build the novel robust model proposed in this paper. In order to solve multi-category classification simultaneously and extend the distance between different classes for a more effective classification. Xiang *et. al* [10] proposed a discriminative LSR(DLSR) model. DLSR makes class label vectors to be the regression targets and introduces a technique called ε-dragging

This work is supported by the National Natural Science Foundation of China (Grant No. 61403376 and 91646207), the Beijing Nature Science Foundation (Grant No. 4162064) and the Open Research Fund of Hunan Provincial Key Laboratory of Network Investigational Technology (Grant No. 2015HNWLFZ055).

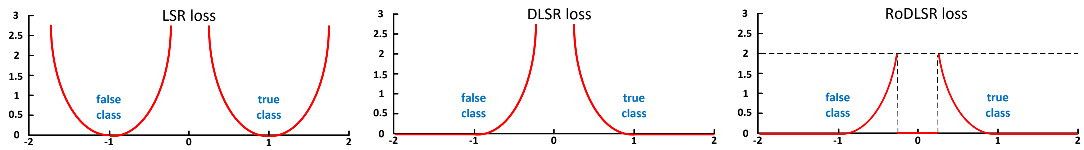

Fig. 1. Comparisons of the loss functions. From left to right are loss functions of LSR, DLSR and the proposed RoDLSR, respectively. For the loss function of RoDLSR, the thresholding value τ is set to be a value of 2.

to enlarge the distance between different class label vectors. That is to say, we assume the right class label as 1 and the wrong class label as -1, and then the ε-dragging transforms the element $+1$ to $1+\varepsilon$, and transforms the element -1 to $-1-\varepsilon$ in the class label vectors.

DLSR achieves better performance than other method in multi-category classification. Due to the compact form and effective solution, some new regression method based on DLSR was suggested. Zhang *et. al* [11] proposed a retargeted LSR (ReDLSR) model and Wang *et. al* [12] proposed margin scalable discriminative LSR (MSDLSR). ReLSR directly learns the regression targets from data and makes the margin between the targets of true and false classes larger than one. So it is much more accurate in measuring the classification error of the regression model than DLSR. In the paper of MSDLSR, the author proved the DLSR is a relaxation of the traditional L_2-support vector machine and the data with dragging values as 0 are the support vector set of DLSR. Based on these discovery, MSDLSR restricts the number of zeros of dragging values by setting a constraint on DLSR to control the margin of DLSR.

ReLSR and MSDLSR both appreciate the superiority of DLSR and refer to the strategy of modifying the regression targets from DLSR, but they neither add the robustness to DLSR. Popular robust regression methods such as M-estimation [13] and LTS [14] are too complex to integrate themselves into the DLSR. So we abandon existing robust algorithm and explore a novel concise robust model called Robust DLSR (RoDLSR) by modifying the regression targets based on DLSR and MSDLSR.

3. METHOD

Given n training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ divided into c categories, where $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ denotes a data point and $y_i \in \{1, 2, \ldots, c\}$ is the class label of $\mathbf{x}_i$. In order to embed the class label information into the formulation and process multi-class classification simultaneously, we denotes a c dimension vector $\mathbf{f}_{y_i}$ as the regression target for the data $\mathbf{x}_i$, where the y_ith element is 1 and the other element is -1. For example, the class label y_i is j for the data point $\mathbf{x}_i$ belonging to the jth category, so the regression target was denoted as

$$\mathbf{f}_{y_i} = [-1, \ldots, -1, 1, -1, \ldots, -1]^T \in \mathbb{R}^c$$

with only the jth element equal to one. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ store the n data points, and $\mathbf{Y} = [\mathbf{f}_{y_1}, \mathbf{f}_{y_2}, \ldots, \mathbf{f}_{y_n}]^T \in \mathbb{R}^{n\times c}$ record their labels. In the following, we first briefly review DLSR model. And then, we describe our RoDLSR model in detail.

3.1. DLSR

Focusing on extending the distance between different categories to enhance the classification accuracy, DLSR trains a ε-dragging matrix attached to the original regression target $\mathbf{Y}$. Assume $\mathbf{B} \in \mathbb{R}^{n\times c}$ be a constant matrix with the ij-th element $\mathbf{B}_{ij}$ defined as

$$\mathbf{B}_{ij} = \begin{cases} +1 & ,\text{if } y_i = j, \\ -1 & ,\text{otherwise} \end{cases}$$

With the constant matrix $\mathbf{B}$, DLSR model is

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{M}} \quad & \|\mathbf{X}\mathbf{W} + \mathbf{e}_n\mathbf{b}^T - \mathbf{Y} - \mathbf{B}\odot\mathbf{M}\|_F^2 + \beta\|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \mathbf{M} \geq 0 \end{aligned} \tag{2}$$

where $\mathbf{W}$ is a transformation matrix in $\mathbb{R}^{d\times c}$, $\mathbf{b}$ is a translation vector in $\mathbb{R}^c$ and $\mathbf{e}_n = [1, 1, \ldots, 1]^T \in \mathbb{R}^n$ is a vector whose all elements are one. $\odot$ is a Hadamard product operator of matrices.

As interpreted in [10], every element of the matrix $\mathbf{B}$ indicates the dragging direction for the corresponding element in matrix $\mathbf{Y}$ and its value is the same as the matrix $\mathbf{Y}$. The nonnegative matrix $\mathbf{M} = \{\varepsilon_{ij} \geq 0\} \in \mathbb{R}^{n\times c}$ records the value of ε-dragging obtained in learning process.Ultimately, the target matrix of DLSR is transformed to $\mathbf{T} = \mathbf{Y} + \mathbf{B}\odot\mathbf{M}$.

3.2. RoDLSR

In [12], Wang et al. have proved that DLSR is a relaxation of L_2-SVM, and the support vector set of DLSR is composed of the samples with dragging values being 0. Hence, DLSR model can select support vectors iteratively by minimizing Eqn. (2). Unfortunately, DLSR does not robust to outlier. To solve this problem, we propose the following RoDLSR model, which is defined as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{M},\mathbf{K}} \quad & \|\mathbf{X}\mathbf{W} + \mathbf{e}_n\mathbf{b}^T - \mathbf{Y} - \mathbf{Y}\odot\mathbf{M} + \mathbf{S}\odot(\tau\mathbf{Y} + \mathbf{Y}\odot\mathbf{K})\|_F^2 \\ & +\beta\|\mathbf{W}\|_F^2 \\ & \text{s.t.} \;\; \mathbf{M} \geq 0, \;\mathbf{K} \geq 0 \end{aligned} \tag{3}$$

where τ is the predefined threshold ($\tau = 2$ used in this paper), which is used to identify data outliers. Matrix $\mathbf{S} \in \mathbb{R}^{n\times c}$ is the selection matrix, and $\mathbf{K} \in \mathbb{R}^{n\times c}$ is the outlier indication matrix, which is also learned in the optimization process. Specifically, the selection matrix $\mathbf{S}$ is defined as

$$\mathbf{S} = \mathbf{C}\odot\mathbf{D}, \tag{4}$$

where $\mathbf{C} \in \mathbb{R}^{n\times c}$ and $\mathbf{D} \in \mathbb{R}^{n\times c}$ are also selection matrices, which are defined as

$$\mathbf{C}_{ij} = \begin{cases} 1, & \mathbf{M}_{i,j} = 0, \\ 0, & \text{otherwise} \end{cases}, \quad \mathbf{D}_{ij} = \begin{cases} 0, & \mathbf{K}_{i,j} = 0, \\ 1, & \text{otherwise} \end{cases}. \tag{5}$$

As illustrated in Eqn. (5), the matrix $\mathbf{C}$ is to judge whether a sample is the support vector (or to select support vectors), while the matrix $\mathbf{D}$ is to judge whether a sample is the outlier (or to select outliers).

Compared with DLSR model, the proposed RoDLSR model introduces the new term $\mathbf{S}\odot(\tau\mathbf{Y} + \mathbf{Y}\odot\mathbf{K})$ for the regression target. In this term, the matrix $\mathbf{S}$ is utilized to select the samples which are support vectors and outliers. After selecting this samples, we add additional label values $\tau\mathbf{Y} + \mathbf{Y}\odot\mathbf{K}$, as a result, the regression losses for these samples are 0. Hence, the proposed RoDLSR model can be robust to the outliers.

3.3. Optimization of RoDLSR Model

The proposed RoDLSR model in Eqn. (3) is a minimized convex quadratic function subjected to two linear constraints. Hence, it is capable to utilize an iterative alternating optimization to acquire the solution with the following three steps.

STEP_1: Given $\mathbf{M}$ and $\mathbf{K}$, and let $\mathbf{R} = \mathbf{Y} + \mathbf{Y} \odot \mathbf{M} - \mathbf{S} \odot (\tau\mathbf{Y} + \mathbf{Y} \odot \mathbf{K}) \in \mathbb{R}^{n\times c}$, the problem of Eqn. (3) can be converted to the classical LSR problem:

$$\min_{\mathbf{W},\mathbf{b}} \|\mathbf{X}\mathbf{W} + \mathbf{e}_n\mathbf{b}^T - \mathbf{R}\|_F^2 + \beta\|\mathbf{W}\|_F^2. \quad (6)$$

The optimal solution can be consulted as

$$\mathbf{W} = (\mathbf{X}^T\mathbf{H}\mathbf{X} + \beta\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{H}\mathbf{R},\ \mathbf{b} = \frac{(\mathbf{R} - \mathbf{X}\mathbf{W})^T\mathbf{e}_n}{n}, \quad (7)$$

where $\mathbf{H} = \mathbf{I}_n - \frac{\mathbf{e}_n\mathbf{e}_n^T}{n}$, $\mathbf{I}_n$ is a $n \times n$ identity matrix and $\mathbf{I}_d$ is a $d \times d$ identity matrix.

STEP_2: Given $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{K}$, the calculation of the matrix $\mathbf{M}$ is similar with that in DLSR model. Denote $\mathbf{P} = \mathbf{X}\mathbf{W} + \mathbf{e}_n\mathbf{b}^T - \mathbf{Y} - \mathbf{S} \odot (\tau\mathbf{Y} + \mathbf{Y} \odot \mathbf{K}) \in \mathbb{R}^{n\times c}$, the solution of $\mathbf{M}$ is

$$\mathbf{M} = \max(\mathbf{Y} \odot \mathbf{P}, 0). \quad (8)$$

The more information about the derivation of Eqn.(8) has been discussed in [10]. Please refer to this paper for details.

STEP_3: Given $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{M}$, and let $\mathbf{Q} = \mathbf{X}\mathbf{W} + \mathbf{e}_n\mathbf{b}^T - \mathbf{Y}$, then, the optimal $\mathbf{K}$ in Eqn. (3) can be calculated by

$$\mathbf{K} = \max((-\mathbf{Y}) \odot \mathbf{Q} - \tau, 0). \quad (9)$$

Note that, with the selection matrix $\mathbf{S}$, we need only consider the support vectors or the samples with $\mathbf{M}_{ij} = 0$.

Proof: Due to the squared Frobenius norm in RoDLSR model in Eqn. (3), we can process the calculation of $\mathbf{K}$ element by element. For the ij-th element $\mathbf{K}_{ij}$ in the outlier matrix $\mathbf{K}$, we have

$$\min_{\mathbf{K}_{ij}} (\mathbf{Q}_{ij} + \tau\mathbf{Y}_{ij} + \mathbf{Y}_{ij}\mathbf{K}_{ij})^2, \quad \text{s.t. } \mathbf{K}_{ij} \geq 0. \quad (10)$$

Since $\mathbf{Y}_{ij}^2 = 1$, we have $(\mathbf{Q}_{ij} + \mathbf{Y}_{ij}\tau + \mathbf{Y}_{ij}\mathbf{K}_{ij})^2 = (\mathbf{Y}_{ij}\mathbf{Q}_{ij} + \tau + \mathbf{K}_{ij})^2$. Combining with the nonnegative constraint of the matrix $\mathbf{K}$ or $\mathbf{K}_{ij} \geq 0$, we can obtain $\mathbf{K}_{ij}$ by

$$\mathbf{K}_{ij} = \max(-\mathbf{Y}_{ij}\mathbf{Q}_{ij} - \tau, 0). \quad (11)$$

Hence, we prove the optimal solution in Eqn. (9).

3.4. Discussion

In this subsection, we further discuss how and why the element $\mathbf{K}_{ij}$ can determine whether a sample is a outlier. As shown in Eqn. (9), we get following conclusions:

- if $\mathbf{K}_{ij} = 0$, and under condition that $\mathbf{Y}_{ij} = 1$ and $\mathbf{M}_{ij} = 0$, we can get that $-\tau < \mathbf{P}_{ij} < 0$.
- if $\mathbf{K}_{ij} = 0$, and under condition that $\mathbf{Y}_{ij} = -1$ and $\mathbf{M}_{ij} = 0$, we can get that $0 < \mathbf{P}_{ij} < \tau$.

To sum up, when $\mathbf{K}_{ij} = 0$ and $\mathbf{M}_{ij} = 0$, the residual of correspond data is within the predefined threshold parameter τ. Therefore, we can select the outliers by $\mathbf{D}$ (related to the matrix $\mathbf{K}$). It worth noting that we can select the support vector set according to $\mathbf{C}$ (related to the matrix $\mathbf{M}$). The comparisons of the loss functions about DLSR and RoDLSR are illustrated in Fig. 1 in detail.

Table 1. Brief description of the data sets.

Info. / Data Set	Classes	Features	Total Num.	Train Num.
Iris	3	4	150	60
Glass	6	9	214	86
SVMGuide2	3	20	391	156
Vehicle	4	18	846	338
Vowel	11	10	990	396
Cora_OS	4	6737(200)	1246	499
Coil20	20	256	1440	576
WebKB-CL	7	4134(200)	827	331
WebKB-WC	7	4189(200)	1210	484

4. EXPERIMENT

In this section, we evaluate the superiority of RoDLSR model by comparing it with the other six multi-category classification models on different error rates for 10 databases. In the following, we first introduce the datasets for evaluation. And then, we explain the selection of parameters in our experiment. Finally, we present the comparable experimental results and analysis.

4.1. Data sets

Nine machine learning databases are utilized in our experiments for evaluating proposed method as [12]. Table 1 introduces the information of these data sets. The first five data sets (Iris, Svmguide2, Vehicle, Glass and Vowel) are downloaded from the LIBSVM machine learning data repository. The Coil20 dataset is used in image classification (face, object, and digit). The last three datasets (Cora-OS, WebKB-CL and WebKB-WC) are widely adopted for information extraction and retrieval.

4.2. Comparison Methods and Parameter Settings

Our method is compared with six multi-category classification methods, including traditional LSR, recent DLSR, L_1-SVM with hinge loss, L_2-SVM with squared hinge loss, logistic regression (LR), and Multiclass SVM (MC-SVM) with multiclass hinge loss. We utilize LIBLINEAR Software [1] to implement the comparative methods, including L_1-SVM, L_2-SVM, LR and MC-SVM. For L_1-SVM, L_2-SVM, LR and MC-SVM, the major regularization parameter C is selected by cross validation, and the candidate set of cross validation is $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2\}$.

From RoDLSR model, shown in Eqn. (3), we need determine two parameters β and τ. The parameter β is defined as

$$\beta = \hat{\beta}\frac{1}{d}tr(\mathbf{X}^T\mathbf{H}\mathbf{X}), \quad (12)$$

where tr$(\cdot)$ is the trace of a matrix, and the parameter $\hat{\beta}$ will be selected by cross validation from $[0, 1]$ whose step is 0.1 to make sure the best possible result. The same approach is utilized to select the regularization parameter β for LSR and DLSR . Moreover, we fix the threshold parameter τ as 2 for all the datasets.

[1] The software is available at www.csie.ntu.edu.tw/∼cjlin/liblinear.

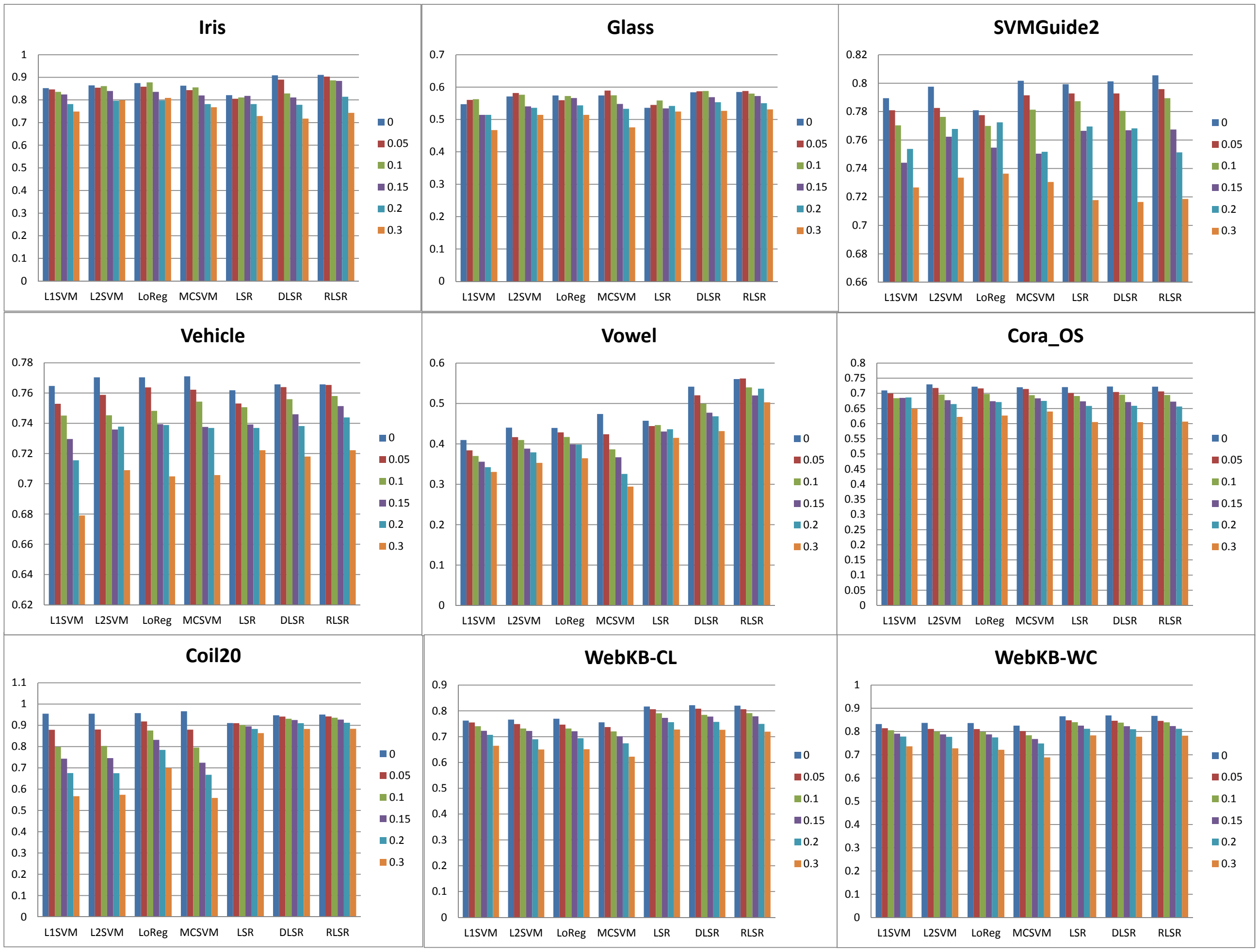

Fig. 2. Comparative results. The title of each subgraph denotes the data set utilized. And the legend of $0, 0.05, 0.1, 0.15, 0.2, 0.3$ in the right side presents the error ratio of the datasets. X axis indicates all the classification method in experiment.

4.3. Results

Based on the determined parameters, we evaluate all the models on the databases. Each dataset is randomly partitioned into the training and testing parts for 10 times (the training data set contains 40% samples, and the rest ones are used for testing). To evaluate the robustness of all models, we pollute $\{0\%, 5\%, 10\%, 15\%, 20\%, 30\%\}$ of the training samples, respectively.

As shown in Fig. 2, it's obvious that the average accuracies of every method for all the datasets descend with the increase of error ratio. Provided any error ratio, the quantity of the data sets getting the best accuracy with RoDLSR model is significantly more than with other methods. Besides, with the increase of error rate, the classification accuracy of RoDLSR model is higher than other SVM-based method for almost data sets, especially Coil20, Glass Vowel and Vehicle, WebKB-WC and WebKB-CL. In these datasets, although the RoDLSR performs inferior to other method under low error ratio, RoDLSR still surpass other methods under the higher error ratio, especially for the datasets Glass, Coil20 and Vehicle.

Compared with DLSR, RoDLSR improves the performances of DLSR, which are significantly obvious in the data sets Iris and Vowel. In these two datasets, the classification accuracies of RoDLSR over DLSR even achieve up to 7%. It's verified that superiority of RoDLSR ascend with the rise of error rate for the data sets Iris, Vehicle and Vowel. These experimental results indicates the effectiveness of RoDLSR for the robust problem.

5. CONCLUSION AND DISCUSSION

In this paper, we propose a simple yet effective RoDLSR model for robust multi-category classification. The core idea behind this model is to control the range of residuals by introducing the outlier term into the original DLSR model. Experimental results evaluate the superiority of RoDLSR in the data sets with outlier samples against the state-of-the-art approaches.

In the future, we will extend RoDLSR model to kernel version by introducing kernel methods [17]. In kernel method, the original feature vector is replaced by the higher dimensional feature vector with nonlinear mapping functions. Same with DLSR, by introducing matrix $\|\cdot\|_{2,1}$ (refer to [10] for details) norm for the regression matrix $\mathbf{W}$, the proposed RoDLSR model can also be used to perform feature selection task, which will be another future work.

6. REFERENCES

[1] S Lemeshow and Hosmer Dw Jr, "A review of goodness of fit statistics for use in the development of logistic regression models.," *American Journal of Epidemiology*, vol. 115, no. 115, pp. 92–106, 1982.

[2] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *J of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[3] Arthur E. Hoerl and Robert W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.

[4] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[5] Christopher J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining And Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[6] J. A. K. Suykens, "Nonlinear modelling and support vector machines," in *IEEE Instrumentation and Measurement Technology Conference*, 2001, pp. 287–294 vol.1.

[7] T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel fisher discriminant analysis," *Neural Computation*, vol. 14, no. 5, pp. 1115–1147, 2002.

[8] Daisuke Tsujinishi and Shigeo Abe, "Fuzzy least squares support vector machines for multiclass problems," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 2, no. 5-6, pp. 785–92, 2003.

[9] Thomas G. Dietterich and Ghulum Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, no. 2, pp. 263–286, 1995.

[10] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection.," *IEEE Transactions on Neural Networks And Learning Systems*, vol. 23, no. 11, pp. 1738–1754, 2012.

[11] X. Y. Zhang, L. Wang, S. Xiang, and C. L. Liu, "Retargeted least squares regression algorithm.," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1, 2014.

[12] L. Wang, X. Y Zhang, and C. Pan, "Msdlsr: Margin scalable discriminative least squares regression for multicategory classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–1, 2015.

[13] Peter J Huber, "Robust regression: asymptotics, conjectures and monte carlo," *The Annals of Statistics*, pp. 799–821, 1973.

[14] Douglas M Hawkins, "The feasible solution algorithm for least trimmed squares regression," *Computational statistics And data analysis*, vol. 17, no. 2, pp. 185–196, 1994.

[15] Philipp Sibbertsen, "S-estimation in the linear regression model with long-memory error terms under trend," *Journal of Time*, vol. 22, no. 3, pp. 353–363, 2001.

[16] Malte Kuss, *Gaussian process models for robust regression, classification, and reinforcement learning*, Ph.D. thesis, TU Darmstadt, 2006.

[17] Changshui Zhang, Feiping Nie, and Shiming Xiang, "A general kernelization framework for learning algorithms based on kernel PCA," *Neurocomputing*, vol. 73, no. 4-6, pp. 959–967, 2010.