Vol.2015-MPS-103 No.55
Vol.2015-BIO-42 No.55
2015/6/25

Kernel Logistic Regression based on the Confusion Matrix for Imbalanced Data Classification

Peng Wang[1] Miho Ohsaki[1] Kenji Matsuda[1] Shigeru Katagiri[1] Hideyuki Watanabe[2]

Abstract: Imbalanced data classification is a common problem in applications related to the detection of anomalies, failures, and risks. Since previous problem-solving approaches were basically heuristic and task dependent, we propose a novel imbalanced data classifier with a theoretical problem-solving approach. Our proposed method fine-tunes the parameters of kernel logistic regression using the harmonic mean of such criteria as sensitivity and positive predictive value, which are derived based on a confusion matrix and are essential for multilateral evaluation. This paper presents the formulation of our proposed method and reports our empirical evaluation results.

1. Introduction

Data that are composed of majority and minority classes are called imbalanced data, which is illustrated in Fig. 1. The classification of imbalanced data is critical for applications related to anomalies, failures, and risks, such as making medical diagnoses and preventing traffic accidents. Conventional methods are categorized into sampling, misclassification costs, or an ensemble of classifiers, but they share a similar approach: the correction of imbalance in a heuristic and task dependent manner [1]. They achieve better results than classifiers with no customization for imbalanced data, but they are less reasonable and less general.

We seek high performance in a theoretically consistent manner and propose an imbalanced data classifier, which we call confusion-matrix-based kernel logistic regression (CM-KLOGR) [2]. CM-KLOGR combines the following techniques, kernel logistic regression (KLOGR) [3], minimum classification error and generalized probabilistic descent (MCE/GPD) learning [4], and evaluation criteria derived from a confusion matrix. It also introduces pretraining and retraining for efficient and effective optimization. We formulated and evaluated our CM-KLOGR and report our results in this paper.

2. CM-KLOGR Formulation

We designed CM-KLOGR to directly and simultaneously raise the values of multifaceted evaluation criteria through smooth optimization. In the beginning, we formulated the posterior probabilities of classes following the manner of KLOGR formulation. In Eq. (1), $\mathbf{x}$ is an input feature vector, $\boldsymbol{\kappa}(\mathbf{x})$ is a kernel function for nonlinear transformation, $\boldsymbol{\alpha}_k$ is a parameter vector to be optimized, and b_k is a cutoff. When we use a Gaussian kernel, its width σ can be a hyperparameter. In Eq. (2), $Pr(C_k|\mathbf{x})$ is the posterior probability of the k-th class given by $\mathbf{x}$, and K is the number of classes. By substituting $y_k(\mathbf{x})$ of Eq.(1) into the softmax function of Eq. (2), $Pr(C_k|\mathbf{x})$ is represented. The cross entropy defined using $Pr(C_k|\mathbf{x})$ is the objective function in CM-KLOGR pretraining.

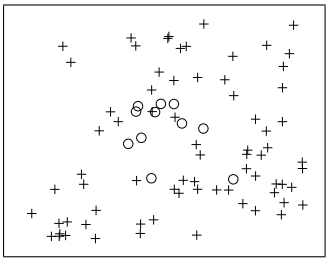

Fig. 1 Imbalanced data consisting of majority and minority classes.

$$y_k(\mathbf{x}) = \boldsymbol{\alpha}_k^T \boldsymbol{\kappa}(\mathbf{x}) + b_k \tag{1}$$

$$Pr(C_k|\mathbf{x}) = \frac{\exp(y_k(\mathbf{x}))}{\sum_{l=1}^{K} \exp(y_l(\mathbf{x}))} \tag{2}$$

For fine-tuning the parameters to improve the multifaceted classification performances, retraining is done using our newly devised objective function. On the basis of MCE/GPD training, we formulated misclassification measure $d_{k_n}(\mathbf{x}_n)$ and its loss function $l(d_{k_n}(\mathbf{x}_n))$ in Eqs. (3) and (4). In these equations, k_n denotes the correct class of the n-th instance, and η is a positive constant. The loss function, Eq. (4), is a differential approximation of a 0-1 loss function that assigns 0 to correct classifications and 1 to incorrect classifications. Here ϵ is a hyperparameter that represents the steepness of the loss function.

$$d_{k_n}(\mathbf{x}_n) = -Pr(C_{k_n}|\mathbf{x}_n) + \left[\frac{1}{K-1} \sum_{j,\ j \neq k_n} Pr(C_j|\mathbf{x}_n)^{\eta} \right]^{\frac{1}{\eta}} \tag{3}$$

$$l(d_{k_n}(\mathbf{x}_n)) = \frac{1}{1 + \exp(-\epsilon d_{k_n}(\mathbf{x}_n))} \quad (\epsilon > 0) \tag{4}$$

[1] Doshisha University
1-3 Tataramiyakodani, Kyotanabe-shi, Kyoto 610-0321, Japan
[2] National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

Now, we can represent several classification patterns based on a confusion matrix. With Eq.(4), we approximated the number of correct classifications for minority class N_{TP}, incorrect classifications for minority N_{FP}, correct ones for majority N_{TN}, and incorrect ones for majority N_{FN}, as shown in Eq. (5).

$$N_{\mathrm{TP}} \approx \sum_{n=1}^{N}(1 - l(d_{k_n}(\mathbf{x}_n)))\delta_{k_n,2}, \quad N_{\mathrm{FP}} \approx \sum_{n=1}^{N} l(d_{k_n}(\mathbf{x}_n))\delta_{k_n,1} \quad (5)$$
$$N_{\mathrm{TN}} \approx \sum_{n=1}^{N}(1 - l(d_{k_n}(\mathbf{x}_n)))\delta_{k_n,1}, \quad N_{\mathrm{FN}} \approx \sum_{n=1}^{N} l(d_{k_n}(\mathbf{x}_n))\delta_{k_n,2}$$

The evaluation criteria are sensitivity (Sens) f_1, positive predictive value (PPV) f_2, specificity (Spec) f_3, negative predictive value (NPV) f_4, and accuracy (Acc) f_5. Their values are obtained by substituting Eq. (5) into their definitions: $f_1 = N_{\mathrm{TP}}/(N_{\mathrm{TP}}+N_{\mathrm{FN}})$, $f_2 = N_{\mathrm{TN}}/(N_{\mathrm{TN}}+N_{\mathrm{FP}})$, $f_3 = N_{\mathrm{TP}}/(N_{\mathrm{TP}}+N_{\mathrm{FP}})$, $f_4 = N_{\mathrm{TN}}/(N_{\mathrm{TN}} + N_{\mathrm{FN}})$, and $f_5 = (N_{\mathrm{TP}} + N_{\mathrm{TN}})/N$.

Finally, our proposed objective function, which is based on the harmonic mean (HM) of these evaluation criteria, is defined in Eq. (6). Here, N_{ec} denotes the number of criteria, γ_i denotes the weight on the i-th criterion, and $\boldsymbol{\mathcal{K}}$ denotes the Gram matrix containing the kernel function values calculated with the training dataset. Note that this HM includes Acc to be able to handle any types of criteria, but theoretically Acc is redundant compared to the other four. If needed to avoid redundancy, it is easily done by setting the weight on Acc to zero (actually we did in our experiments).

$$J = -\left[\frac{\sum_{i=1}^{N_{\mathrm{ec}}}\gamma_i\frac{1}{f_i}}{\sum_{i=1}^{N_{\mathrm{ec}}}\gamma_i}\right]^{-1} + \frac{\lambda}{2}\sum_{k=1}^{K}\boldsymbol{\alpha}_k^T\boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_k \qquad (6)$$

The first term of Eq.(6) is the harmonic mean of criteria, namely HM, and directly raises their values. The second term is the L2 norm of the parameters with weight λ, which is a hyperparameter, for the suppression of overfitting. CM-KLOGR simultaneously improves various types of performances, avoiding overfitting through smooth optimization due to the initialized parameter setting in the pretraining.

3. CM-KLOGR Evaluation

We conducted an experiment to comparatively evaluate the CM-KLOGR performances with two competitive classifiers. One is kernel logistic regression (KLOGR), which is the basis of CM-KLOGR that leads to an almost equivalent outcome as that of CM-KLOGR's pretraining. The other is support vector machine (SVM), which is an effective widespread kernel method that possesses a different loss definition than ours.

We used the imbalanced datasets in Table 1 [5]. The 10% instances were picked out of a dataset for a test, and the remaining 90% were used for training and validation. The remaining set was divided into ten subsets for 10-fold cross-validation and used to set the values of the parameters, the hyperparameters, and the cutoff. Under their best setting, we estimated the generalized performances using the test set, which were represented by HM and its elements (Sens, Spec, PPV, NPV, and Acc).

Table 1 Specifications of benchmark datasets, where Maj. and Min. denote majority and minority classes.

Name of Datasets	Number of Features	Size of Maj./Min. (Total)	Ratio of Maj. to Min.
Haberman's Survival	3	225/81 (306)	2.78
Pima Indian Diabetes	8	500/268 (768)	1.87

Table 2 Classification performances obtained in experiment.

Haberman's Survival Dataset

	Sens	Spec	PPV	NPV	Acc	HM
CM-KLOGR	50.00	100.00	100.00	85.19	87.10	77.31
KLOGR	62.50	78.26	50.00	85.71	74.19	66.18
SVM	75.00	78.26	54.55	90.00	77.42	72.00

Pima Indian Diabetes Dataset

	Sens	Spec	PPV	NPV	Acc	HM
CM-KLOGR	70.37	84.00	70.37	84.00	79.22	76.58
KLOGR	70.37	82.00	67.86	83.67	77.92	75.34
SVM	81.48	70.00	59.46	87.50	74.03	72.99

Table 2 provides the experimental results. For the Haberman's Survival Dataset in the upper section, CM-KLOGR outperformed KLOGR and SVM for many of the evaluation criteria. CM-KLOGR's comprehensive HM performance exceeds that of KLOGR at a level of 11.13%. Compared to SVM, CM-KLOGR's HM was higher by 5.31%. A similar trend appears for the Pima Indian Diabetes Dataset in the lower section of Table 2. It was suggested by these results that CM-KLOGR is superior to KLOGR and SVM.

If an application requires higher Sens, we can set the weight on Sens larger in CM-KLOGR and so will try this setting in the future. It also will be our future work to conduct further experiments under more finely adjusted hyperparameters.

4. Conclusions

We proposed an imbalanced data classifier, called confusion-matrix-based kernel logistic regression (CM-KLOGR). CM-KLOGR has a theoretical mechanism to raise the values of all the evaluation criteria, including sensitivity, positive predictive value, and so on with no heuristics or task dependent procedures. It outperformed kernel logistic regression and support vector machine under our experimental conditions.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 15K00323 and a MEXT-supported Program for the Strategic Research Foundation at Private Universities 2014-2018.

References

[1] He, H. and Garcia, E. A.: Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, (2009).

[2] Ohsaki, M., Matsuda, K., Wang, P., Katagiri, S., and Watanabe, H.: Formulation of the Kernel Logistic Regression based on the Confusion Matrix, *IEEE Congress on Evolutionary Computation CEC-2015*, accepted, (2015).

[3] Roth, V.: Probabilistic Discriminative Kernel Classifiers for Multi-class Problems, *Lecture Notes in Computer Science*, vol. 2191, pp. 246–253, (2001).

[4] Juang, B. H. and Katagiri, S.: Discriminative Learning for Minimum Error Classification, *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, (1992).

[5] Lichman, M.: UCI Machine Learning Repository (online), available from ⟨https://archive.ics.uci.edu/ml/datasets.html⟩ (accessed 2015-04-28).