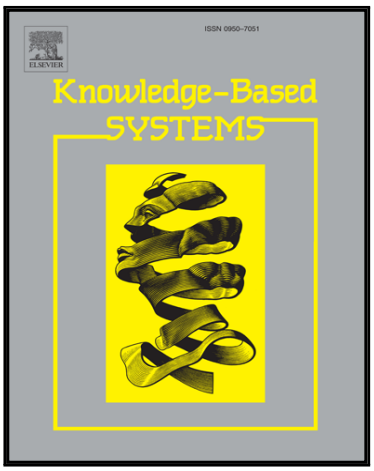

Regularized Fisher Linear Discriminant through two threshold variation strategies for imbalanced problems

Yujin Zhu, Zhe Wang, Chenjie Cao, Daqi Gao

PII: S0950-7051(18)30098-4
DOI: 10.1016/j.knosys.2018.02.035
Reference: KNOSYS 4243

To appear in: *Knowledge-Based Systems*

Received date: 17 May 2017
Revised date: 21 February 2018
Accepted date: 24 February 2018

Please cite this article as: Yujin Zhu, Zhe Wang, Chenjie Cao, Daqi Gao, Regularized Fisher Linear Discriminant through two threshold variation strategies for imbalanced problems, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.02.035

Regularized Fisher Linear Discriminant through two threshold variation strategies for imbalanced problems

Yujin Zhu[1,2], Zhe Wang[*,1,2], Chenjie Cao[2], Daqi Gao[*,2]

Abstract

Fisher Linear Discriminant (FLD) has been widely applied to classification tasks due to its simple structure, analytical optimization, and useful criterion. However, when dealing with imbalanced datasets, even though the weight vector of FLD could be trained correctly to preserve the global distribution information of samples, the threshold of FLD might be seriously misled by the extreme proportion of classes. In order to modify the threshold and preserve the weight vector at the same time so as to improve FLD in imbalanced cases, this paper first regularizes the original FLD in a way inspired by the locality preserving projection, and then utilizes two strategies to optimize the threshold: the multi-thresholds selection strategy trains several FLDs with different empirically-defined thresholds, and then selects the optimal threshold out; the threshold-eliminated strategy generates two hyperplanes parallel to the original one built by FLD, and then utilizes a heuristic similarity metric for prediction. It is seen that the former seeks new threshold instead of the old one, while the latter ignores the original threshold. After introducing both strategies into the regularized FLD, two new classifiers are proposed in this paper and abbreviated as RFLD-S1 and RFLD-S2, respectively. Subsequently, the comprehensive comparison experiments on forty-one datasets among nine typical classifiers validate the effectiveness of the proposed methods. Especially, RFLD-S1 performs better than RFLD-S2 and achieves the best on most datasets.

Index Terms

[*] Corresponding authors. Email: {wangzhe, gaodaqi}@ecust.edu.cn.

[1] Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai, 200237, P.R. China.

[2] Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, P.R. China.