Available online at www.sciencedirect.com

ScienceDirect

Journal of the Franklin Institute 000 (2017) 1–19

www.elsevier.com/locate/jfranklin

Exploiting the categorical reliability difference for binary classification

Lei Sun[a], Kar-Ann Toh[b,*], Badong Chen[c], Zhiping Lin[d]

[a] School of Information and Electronics, Beijing Institute of Technology Beijing, Beijing 100081, PR China
[b] School of Electrical and Electronic Engineering, Yonsei University Seoul, Seoul 120–749, Republic of Korea
[c] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, PR China
[d] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

Received 1 December 2016; received in revised form 24 October 2017; accepted 11 November 2017
Available online xxx

Abstract

In binary pattern classification, the reliabilities of statistics obtained from the samples of the two categories are generally different. When the statistics are used for modeling a classifier, such reliability difference could impact the generalization performance. We formulate a disparity index to show the statistical disparity based on the generalized eigenvalue decomposition of the categorical moment matrices. It is shown that this disparity index can effectively indicate the reliability difference between the two categories. The obtained reliability difference is subsequently utilized to adjust the regularization term of a classifier for effective learning generalization. Our experiments based on 10 real-world benchmark data sets validate the effectiveness of the proposed method.

1. Introduction

In pattern classification, a common attempt of most learning processes is to capture relevant discriminative information based on the statistics acquired from the given pool of training samples. The categorical statistics acquired from the training samples are subsequently used to predict the labels of the testing samples. Therefore, the acquired statistics are expected

* Corresponding author.
E-mail addresses: sunlei@bit.edu.cn (L. Sun), katoh@ieee.org (K.-A. Toh), chenbd@mail.xjtu.edu.cn (B. Chen), ezplin@ntu.edu.sg (Z. Lin).

https://doi.org/10.1016/j.jfranklin.2017.11.024
0016-0032/

to be consistent, or reliable, between the training samples and the testing samples. However, the reliabilities of the acquired statistics among the categories can generally be different. The categorical reliability difference in statistics arouses research interests to come up with a reliable representation for good generalization performance [1–3]. As an example from the literature, under the setting of a binary Bayesian classifier, a common covariance matrix for the two categories has been considered [2, see (77) in Section 3.7 on page 32]. To reflect the effects of the covariance matrices, an affine combination, instead of an equal treatment, of the estimated class-dependent covariance matrices has been used. Effectively, the affine combination adjusts the significance of the two class-dependent covariance matrices to obtain a reliable representation for the two categories. This example shows the existence of the reliability difference between the two categories for binary classification.

In order to utilize the statistical reliability difference, the first step is to know the reliability difference between the two categories. If there is no prior knowledge about the properties of the two categories, it is difficult to determine which category is more reliable. The number of categorical samples can provide certain prior knowledge about the categorical properties. For example, in the binary linear classification model based on the least squares [4, Section 4.1.3 on p.184], its solution implicitly assumes that the statistics obtained from the category with a larger sample number is more reliable. Another example is the Asymmetric Principle Component Analysis classifier (APCA) [5] where the class-dependent covariance matrices are weighted by utilizing the prior knowledge from the number of samples. While the number of samples can provide certain prior knowledge for the weighting procedure, it is reasonable to ask whether the weighted matrix is more reliable than the original one and how to indicate its reliability.

The question regarding the indication of categorical reliability is to be studied in this work. Subsequently, we study exploiting the categorical reliabilities to improve the generalization performance for binary classification. In particular, a disparity index is proposed to indicate the reliability difference between the class-dependent moment/covariance matrices. The disparity index is obtained from the Generalized Eigen-value Decomposition (GED) [6] of the moment/covariance matrices. Based on this disparity index, the reliability difference between the two categories is utilized by suppressing those unreliable components of the moment/covariance matrices in order to improve the generalization performance. We focus on a linear classifier which uses the Total Error Rate classifier (TER) [7] as the training objective for evaluating the proposed method.

Our contributions are listed as follows. (i) A quantitative disparity index is formulated to evaluate the reliability of the class-dependent matrices for binary classification. This index is defined as the disparity between the two GED diagonalized matrices. (ii) The generalization performance of a representative classifier, the TER classifier, is studied by analyzing the reliability of its moment matrices. It is found that a pure re-weighting of the class-dependent matrices does not significantly improve the generalization performance. (iii) A penalized TER classification algorithm is proposed based on a combination of the re-weighted matrix with an additional penalty term. (iv) The effectiveness of the proposed algorithm is validated based on 10 real-world benchmark data sets.

The rest of this paper is organized as follows. In Section 2, a disparity index is proposed to indicate the reliability difference between the class-dependent matrices. In Section 3, a penalized TER classification algorithm is proposed by exploiting the reliability difference between the binary categories. The proposed algorithm is validated by real-world data sets in Section 4. Conclusion is finally given in Section 5.

Please cite this article as: L. Sun et al., Exploiting the categorical reliability difference for binary classification, Journal of the Franklin Institute (2017), https://doi.org/10.1016/j.jfranklin.2017.11.024