Paper Number: 2519

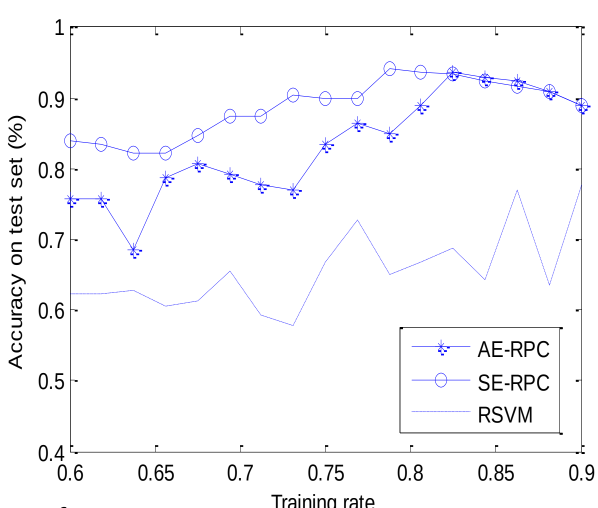
A General Robust Probability Classifier Based on the φ-divergences and Its Application on Lithology Classification

Wang, Y. [1], Zhang, Y. [2] and Cheng, Q. [3,4]

[1]College of Instrumentation & Electrical Engineering, Jilin University, Changchun 130061, China (iamwangyongzhi@126.com,corresponding author)
[2]Department of Industrial Engineering, Tsinghua University, Beijing 100084, China (zhangyuili@mail.tsinghua.edu.cn)
[3]State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing 100083,Wuhan 430074,China
[4]Earth and Space Science and Engineering, York University,Toronto,M3J1P3, Canada.(qiuming@yorku.ca)

In this paper, we propose a general robust probability classifier (RPC) based on φ-divergences to address classification problems with data uncertainty. Traditional classifiers make an exact distributional assumption about the class-conditional densities, which may be unavailable due to unavoidable observational noises and other uncertain factors. To address these issues, we propose a class-conditional probability distributional set based on φ-divergences to describe the data uncertainty. The optimal RPC is defined as the one with the minimal worst-case loss function value over all possible distributions in the proposed distributional set. For RPC models under an absolute error loss criterion (AE-RPC), we give general equivalent reformulations and show that the corresponding problems can be solved as polynomial-time-solvable second order cone programming or linear programming problems. For RPC models under a squared error loss criterion (SE-RPC), we define a modified distributional set based on the modified χ^2-distance and show that it can be solved as an equivalent semi-definite programming problem.

Some experiments on binary and multiple classification problems validate that the proposed models are robust to the data uncertainty and can avoid the “over-learning” phenomenon.



Experimental results show that the proposed RPC models outperform RSVM for both binary and multiple classification problems on the tested data sets. For example, Figure 1 shows the performance of AE-RPC, LE-RPC and RSVM on Y5 data set when different training rates are selected. In general, SE-RPC provides the highest classification accuracy on the test sets and AE-RPC provides the best classification accuracy on the training sets among these models. Both AE-RPC and SE-RPC have the robustness to the data uncertainty and can avoid the “over-learning” phenomenon.

Figure 1: Performance of AE-RPC, SE-RPC and RSVM on Y5 data set.

References:

[1] Huang K (2004) The Journal of Machine Learning Research 5:1253-1286

[2] Kitahara T (2008) Journal of the Operations Research Society of Japan 51(2):191-201
[3] Delage E (2010) Operations Research 58(3):595-612.
[4] Bental A (2013) Management Science 59(2):341-357
[5] Wang Y and Zhang Y (2014) Mathematical Problems in Engineering 2014:1-13