METRIC OPTIMIZATION IN SEMISUPERVISED LEARNING PROBLEMS[1]

G. Iofina, A. Minaev, Yu. Polyakov, Yu. Maximov

Moscow Institute of Physics and Technology, Moscow Region
e-mail: giofina@mail.ru

We consider semi-supervised learning problems in non-transductive setting, e.g. given a set of l labeled examples and $u \gg l$ unlabeled we need to predict classification of unlabeled (and probably unseen) set with high accuracy. Here we assume, that all classes contains at least one labeled object.

This problem arises in a wide variety of applications starting from medical diagnostics and image annotation to modern informational retrieval problems. One of the standard approaches to attack this problems consists of 2 stages. First, perform classification in area of high confidence. After that, algorithms classifies the rest part of the sample based on the information gathered on the first stage [4,5].

One consider within the paper SLL methods follow the schema above. Following [4] we use nearest neighbors method on the first stage. Classical obstacle here is exponential dependence of minimal number of labeled objects required for tight classification. Our contribution is generalization the method from [5] for multiclass problems. Also we reduce the bound for the minimal number of labeled examples required for discrete metric sets w.r.t. to the results proposed in [5] adapting the ideas proposed in [1–3].

REFERENCES

1. N. Goyal, Y. Lifshits,H. Schütze. *Disorder inequality: a combinatorial approach to nearest neighbor search.* — WSDM '08 Proceedings of the International Conference on Web Search and Data Mining. — 2008. P. 25–32.
2. G.V. Iofina. *Optimal metrics in classification problems with ordered features and an arbitrary number of classes.* — Pattern Recognition and Image Analysis. — 2009. Vol. 19. Iss. 2. P. 284–288
3. G.V. Iofina. *A study of metrics in finite sets for application in classification and recognition problems.* — Computational Mathematics and Mathematical Physics. — 2010. Vol. 50. Iss. 3, P. 558–565
4. P. Rigollet. *Generalization Error Bounds in Semi-supervised Classification Under the Cluster Assumption.* — Journal of Machine Learning Research. — 2007. Vol. 8. P. 1369–1392.
5. R. Urner, S. Wulff, S. Ben-David. *PLAL: Cluster-based active learning.* — Journal of Machine Learning Research. Workshop and Conference Proceedings. — 2013. Vol. 30. P. 376–397.

[1] The reported study was supported by RFBR, research project No. 14-07-31241 мол_a and No. 14-07-31241 мол_a; work is also supported in part by Laboratory for Structural Methods of Data. Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.