Cai, Ruichu; **Zhang, Zhenjie**; **Hao, Zhifeng**

BASSUM: a Bayesian semi-supervised method for classification feature selection. (English)

Zbl 1213.68517
Pattern Recognition 44, No. 4, 811-820 (2011).

Summary: Feature selection is an important preprocessing step for building efficient, generalizable and interpretable classifiers on high dimensional data sets. Given the assumption on the sufficient labelled samples, the Markov Blanket provides a complete and sound solution to the selection of optimal features, by exploring the conditional independence relationships among the features. In real-world applications, unfortunately, it is usually easy to get unlabelled samples, but expensive to obtain the corresponding accurate labels on the samples. This leads to the potential waste of valuable classification information buried in unlabelled samples.

In this paper, we propose a new BAyesian Semi-SUpervised Method, or BASSUM in short, to exploit the values of unlabelled samples on classification feature selection problem. Generally speaking, the inclusion of unlabelled samples helps the feature selection algorithm on (1) pinpointing more specific conditional independence tests involving fewer variable features and (2) improving the robustness of individual conditional independence tests with additional statistical information. Our experimental results show that BASSUM enhances the efficiency of traditional feature selection methods and overcomes the difficulties on redundant features in existing semi-supervised solutions.

MSC:

68T10 Pattern recognition, speech recognition
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

feature selection; semi-supervised; structured object; Markov blanket; conditional independence test

Software:

TETRAD; TMG

Full Text: DOI

References:

[1] $\langle$http://discover1.mc.vanderbilt.edu/discover/public/causal_explorer/index.html$\rangle$.

[2] $\langle$http://pages.cs.wisc.edu/⊠dpage/kddcup2001/$\rangle$.

[3] $\langle$http://people.csail.mit.edu/jrennie/20newsgroups/$\rangle$.

[4] $\langle$http://www.causality.inf.ethz.ch/data/sido.html$\rangle$.

[5] $\langle$http://www.csie.ntu.edu.tw/⊠cjlin/libsvm/$\rangle$.

[6] Aliferis, C.; Tsamardinos, I.; Statnikov, A., Hiton, a novel Markov blanket algorithm for optimal variable selection, (), 21-25

[7] Cooper, G.F.; Herskovits, E., A Bayesian method for the induction of probabilistic networks from data, Machine learning, 9, 309-347, (1992) · Zbl 0766.68109

[8] Dietterich, T.; Domingos, P.; Getoor, L.; Muggleton, S.; Tadepalli, P., Structured machine learning: the next ten years, Machine learning, 73, 1, 3-23, (2008)

[9] Dougherty, J.; Kohavi, R.; Sahami, M., Supervised and unsupervised discretization of continuous features, (), 194-202

[10] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V., Gene selection for cancer classification using support vector machines, Machine learning, 46, 1, 389-422, (2002) · Zbl 0998.68111

[11] Handl, J.; Knowles, J., Semi-supervised feature selection via multiobjective optimization, ()

[12] Kohavi, R.; John, G.H., Wrappers for feature subset selection, Artificial intelligence, 97, 1-2, 273-324, (1997) · Zbl 0904.68143

[13] Koller, D.; Sahami, M., Toward optimal feature selection, ()

[14] Liu, H.; Sun, J.; Liu, L.; Zhang, H., Feature selection with dynamic mutual information, Pattern recognition, 42, 7, 1330-1339,

(2009) · Zbl 1183.68540

[15] Margaritis, D.; Thrun, S., Bayesian network induction via local neighborhoods, (), 505-511

[16] Mitchell, T.M., Machine learning, (1997), McGraw-Hill · Zbl 0913.68167

[17] Ramaswamy, S.; Tamayo, P.; Rifkin, R.; Mukherjee, S.; Yeang, C.; Angelo, M.; Ladd, C.; Reich, M.; Latulippe, E.; Mesirov, J.P.; Poggio, T.; Gerald, W.; Loda, M.; Lander, E.S.; Golub, T.R., Multiclass cancer diagnosis using tumor gene expression signatures, Proceedings of the national Academy of sciences of the united states of America, 98, 26, 15149-15154, (2001)

[18] Spirtes, P.; Glymour, C.; Scheines, R., Causation, prediction, and search, (2001), The MIT Press · Zbl 0981.62001

[19] I. Tsamardinos, C.F. Aliferis, A. Statnikov, Time and sample efficient discovery of Markov blankets and direct causal relations, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 673-678.

[20] I. Tsamardinos, C.F. Aliferis, A.R. Statnikov, Algorithms for large scale Markov blanket discovery, in: Proceedings of FLAIRS Conference, 2003, pp. 376-381.

[21] Yang, J.J.; Yang, M.C., An improved procedure for gene selection from microarray experiments using false discovery rate criterion, BMC bioinformatics, 7, 15, (2006)

[22] D. Zeimpekis, E. Gallopoulos, Tmg: a matlab toolbox for generating term-document matrices from text collections, Technical Report, University of Patras, 2005.

[23] Z. Zhao, H. Liu, Semi-supervised feature selection via spectral analysis, in: Proceedings of SIAM International Conference on Data Mining, SIAM, 2007.

[24] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of International Conference on Machine Learning, ACM, 2007, pp. 1151-1157.

[25] Zhou, X.; Mao, K.Z., LS bound based gene selection for DNA microarray data, Bioinformatics, 21, 8, 1559-1564, (2005)