Abstract

In this paper a hybrid feature selection method is proposed which takes advantages of wrapper subset evaluation with a lower cost and improves the performance of a group of classifiers. The method uses combination of sample domain filtering and resampling to refine the sample domain and two feature subset evaluation methods to select reliable features. This method utilizes both feature space and sample domain in two phases. The first phase filters and resamples the sample domain and the second phase adopts a hybrid procedure by information gain, wrapper subset evaluation and genetic search to find the optimal feature space. Experiments carried out on different types of datasets from UCI Repository of Machine Learning databases and the results show a rise in the average performance of five classifiers (Naive Bayes, Logistic, Multilayer Perceptron, Best First Decision Tree and JRIP) simultaneously and the classification error for these classifiers decreases considerably. The experiments also show that this method outperforms other feature selection methods with a lower cost.
- Silberschatz, A., Korth, H. F., and Sudarshan, S., 2010. Database System Concepts, McGrawHill.
- Bellman, R., 1996. Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton.
- Ladha, L., Deepa, T., 2011. Feature Selection Methods and Algorithms, International Journal on Computer Science and Engineering, Volume 3, page(s). 1787-1790.
- Liu, H., and Zhao, Z., 2012. Manipulating Data and Dimension Reduction Methods: Feature Selection, Journal of Computational Complexity, Page(s) 1790-1800.
- Janecek, G. K., Gansterer, N., Demel, A., and Ecker, F., 2008. On the relationship between feature selection and classification accuracy, Journal of Machine Learning and Research. JMLR: Workshop and Conference Proceedings 4, Pages 90–105.
- Assareh, A., Moradi, M., and Volkert, L., 2008. A hybrid random subspace classifier fusion approach for protein mass spectra classification, Springer, LNCS, Volume 4973, Page(s) 1–11, Heidelberg.
- Hayward, J., Alvarez, S., Ruiz, C., Sullivan, M., Tseng, J., and Whalen, G., 2008. Knowledge discovery in clinical performance of cancer patients, IEEE International Conference on Bioinformatics and Biomedicine, USA, Page(s) 51–58.
- Dhiraj, K., Santanu Rath, K., and Pandey, A., 2009. Gene Expression Analysis Using Clustering, 3rd international Conference on Bioinformatics and Biomedical Engineering.
- Jiang, B., Ding, X., Ma, L., He, Y., Wang, T., and Xie, W., 2008. A Hybrid Feature Selection Algorithm: Combination of Symmetrical Uncertainty and Genetic Algorithms, The Second International Symposium on Optimization and Systems Biology, Page(s) 152–157, Lijiang, China, October 31– November 3.
- Zhou, J., Peng, H., and Suen, C., 2008. Data-driven decomposition for multi-class classification, Journal of Pattern Recognition, Volume 41, Page(s) 67 – 76.
- Fayyad U., Piatetsky-Shapiro, G., and Smyth, P., 1996. From Data Mining A Knowledge Discovery in Databases, American Association for Artificial Intelligence.
- Novakovic, J., 2010. The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier, 18th Telecommunications Forum TELFOR, page(s) 1114-1116, November 23-25, Serbia, Belgrade.
- Domingos, P., Pazzani, M., 1997. On the Optimality of the Simple Bayesian Classifier under Zero-One loss, Machine Learning, Volume29, page(s) 103-130, November/December 1997.
- Blum, A. L., Rivest, R. L., 1992. Training a 3-node neural networks is NP-complete, Neural Networks, Volume 5, page(s) 117-127.
- Almuallim, H., Dietterich, T. G., 1991. Learning with many irrelevant features, in proceedings of AAAI-91, page(s) 547-552, Anaheim, California.
- Blum, A. L., Langley, P., 1997. Selection of Relevant Features and Examples in Machine Learning, Artificial Intelligence, Volume 97, page(s) 245-271.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Volume 16, page(s) 321-357.
- Kohavi, R., John, G. H., 1997. Wrappers for Feature Subset Selection, Artificial Intelligence, Volume 97, page(s) 273-324.
- Olusola, A. A., Oladele, A. S., Abosede, D. O., 2010. Analysis of KDD'99
Intrusion Detection Dataset for Selection of Relevance Features, Proceedings of the World Congress on Engineering and Computer Science, October 20-22, San Francisco, USA.
- Randy, H., and Haupt, S., 1998. Practical Genetic Algorithms, John Wiley and Sons.
- Mertz C. J., and Murphy, P. M., 2013. UCI Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California.
- Dash, M., Liu, H., 2003. Consistency-based Search in Feature Selection, Artificial Intelligence, Volume 151, page(s) 155-176.
- Chou, T. S., Yen, K. K., and Luo, J., 2008. Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms, International Journal of Information and Mathematical Sciences, Volume 4, page(s) 196-208.

Index Terms

Computer Science
Data Mining

Keywords
Feature Selection Resampling Information Gain Wrapper Subset Evaluation