Abstract

The fundamental population data are needed for every country for purposes of planning, development, and improvement. Census data can provide the basic population data of any country. Moreover, they are rich with lots of hidden information that can be used for machine learning and data mining tasks in order to provide services for country's social and economic development. This paper is focused on the applications of data mining and machine learning in census data to classify the annual income. It aims to show a systematic comparison to examine and evaluate three supervised learning classifiers. The classifiers that have been targeted are decision trees, random forests, and artificial neural networks. The main aims are to explore not only the classifiers properties and the impact of the attributes on the evaluation, but also, evaluate their classification performance under certain conditions to understand how the performance of the models changes over different experiments which potentially provide a guidance to help researchers to determine the most suitable classifier in census data.
1. Chikohora, T. (2014). A Study of The Factors Considered When Choosing an Appropriate Data Mining Algorithm. International Journal of Soft Computing and Engineering (IJSCE), 4(3), 1-6.
2. Sumathi, S., & Sivanandam, S. (2006). Introduction to Data Mining and its Applications. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-540-34351-6
3. Hassani, H., Saporta, G., & Silva, E. (2014). DATA MINING AND OFFICIAL STATISTICS: The Past, the Present and the Future. The journal of big data, 2(1), 34-43. doi:10.1089/big.2013.0038
4. Wheldon MC, Raftery AE, Clark SJ, Gerland P. Estimating Demographic Parameters with Uncertainty from Fragmentary Data. Center for Statistics and the Social Sciences, University of Washington, Seattle, Washington, 2011, Working Paper 108.
5. Drummond, C., Matwinm, S., & Gaffield, C. (2000). Inferring and revising theories with confidence: data mining the 1901 Canadian census. Journal of Machine Learning Research, 1-48. doi:10.1080/08839510500313711
6. Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. Journal of Official Statistics, 12(4), 385-401.
7. Nithya, A., & Sundaram, V. (2011). Wheat disease identification using Classification Rules. International Journal of Scientific & Engineering Research, 2(9), 01-05.
8. Malerba, D., Esposito, & Lisi, F. (2002). Mining Spatial Association Rules in Census Data. Intelligent Data Analysis, 541-550.
9. Han, J., & Kamber, M. (2006). Data Mining: Concepts and Technique (2nd ed.). Morgan Kaufmann.
10. Witten, I., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.). Morgan Kaufmann.
11. Breiman, L. (1996). RANDOM FORESTS. Machine Learning journal, 45(1), 1-33.
12. Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272-278.
13. Williams, J., Ahijevych, D. K., Saxen, T., Steiner, M., & Dettling, S. (2008). A machine-learning approach to finding weather regimes and skillful predictor combinations for short-term storm forecasting. American Meteorological Society Journal (AMS), 1-6.
14. Berry, M., & Linoff, G. (2004). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (2nd ed.). Wiley.
15. Alon, I., Qi, M., & Sadowski, R. (2001). Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. Journal of Retailing and Consumer Services, 8(3), 147-156. doi:10.1016/S0969-6989(00)00011-4

Index Terms

Computer Science
Algorithms
Keywords

Census Data, Data Mining, Classification, Supervised Learning, Decision Trees, Random Forests, Artificial Neural Networks, Performance Metrics.