A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data

Abstract

Data Mining (the analysis step of the Knowledge Discovery in Databases process or KDD), a relatively young and interdisciplinary field of computer science, is the process of discovering or
extracting new patterns from large data sets involving methods from statistics and artificial intelligence. It is commonly used in marketing, surveillance, fraud detection, scientific discovery and now gaining wide way in social networking. Anything and everything on the Internet is fair game for extreme data mining practices. Social media covers all aspects of the social side of the internet that allow us to get contact and carve up information with others as well as intermingle with any number of people in any place in the world. This paper uses the dataset “Social side of the Internet” from Pew Research Center. The focus of the research is towards exploration on impact of the internet on social group activities using Data Mining Techniques. The original dataset contains 162 attributes which is very large and hence the essential attributes required for the analysis are selected by feature reduction method. The selected attributes were applied to Data Mining Classification Algorithms such as RndTree, ID3, K-NN, C-RT, CS-CRT, C4.5 and CS-MC4. The Error rates of various classification Algorithms were compared to bring out the best and effective Algorithm suitable for this dataset.

Reference

- Tanagra Data Mining tutorials, http://data-mining-tutorials.blogspot.com/ This website provides detailed information on the basics of Data Mining Algorithms
- Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 2008-12-17.
Index Terms

Computer Science

Pattern Recognition

Key words

Knowledge discovery in databases
data mining

surveys