Abstract

A spam has diluted the message pool, causing frustration so require an automatic processing of emails. This study is to construct a spam model using classification technique in data mining. To accomplish this, experiments were conducted on spam dataset downloaded from the UCI machine learning repository which was classified using a popular data mining tool called WEKA. The final classification result should be 1 if it is finally spam, otherwise, it should be 0. Email is popular mode of communication and its users are growing day by day. But, due to social networks and electronic business, most of the emails contain unsolicited bulk e-mail called spam. Several solutions have been proposed to overcome the spam problem, filtering using decision tree classifiers is the one of the most significant techniques. Machine learning classifiers, J48, J48graft and Simple CART were used for classifying spam messages from e-mail. These trees are induced first and then prune sub trees to improve classification accuracy and size of tree. It helps to reduce size, complexity and to achieve better predictive accuracy of final classifier. Grafting is then applied as a post process to an inferred decision tree. Results showed that J48graft had pretty good prediction accuracy as compared to CART and J48 algorithms.
References

- J. Quinlan, Simplifying decision trees, Int. J. Human Computer Studies.
- SamDrazin and MattMontag, Decision Tree Analysis using WEKA, Machine Learning-Project II, University of Miami.
- Max Bramer, "Pre-pruning Classification Trees to reduce Overfitting in Noisy Domains"; Faculty of Technology, University of Portsmouth, UK.

Index Terms

Computer Science  Artificial Intelligence

Keywords

Weka  Simple CART  J48  J48graft  Spam filtration  Post pruning  Pre pruning

Classification

Grafting