Abstract

Feature selection has been a productive field of research and development in data mining, machine learning and statistical pattern recognition, and is widely applied to many fields such as, image retrieval, genomic analysis and text categorization. Feature selection includes selecting the most useful features from the given data set. The feature selection involves removing irrelevant and redundant features form the data set. The feature selection can be efficient and effective using clustering approach. Based on the criteria of efficiency in terms of time complexity and effectiveness in terms of quality of data, useful features from the big data can be selected. Feature selection reduces the computational complexity of learning and prediction algorithms and saves on the cost of measuring non selected features. The feature selection can be done using the graph clustering approach based on theoretic graph. The most
relevant features are selected from the cluster for the relevant target class. The features in every cluster are different and independent of the other.

References

- Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE transaction on Knowledge and Data Engineering 2013
- Yu L. and Liu H., "Redundancy based feature selection for microarray data", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004
- Pereira F., Tishby N. and Lee L., "Distributional clustering of English Words", Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, pp
183-190, 1993.

Index Terms
Computer Science
Database Management

System

Keywords
Feature Selection  Clustering