Part Pramokchon 2014: Content-Adaptive Filter based Feature Selection. Doctor of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Punpiti Piamsa-nga, D.Sc. 92 pages.

In data representation approaches, a feature vector model using only some relevancy features, also known as optimal feature set, instead of all features yields better data analysis performance. Filter-based method is the most used feature selection method for highly dimensional data. However, there are three important drawbacks of filter-based methods. First, the optimal set depends on empirical based algorithm. Second, the ranking method ignore considering the redundancy between features which reduces the overall performance. Finally, feature scoring used for ranking feature often fails in the imbalanced class distribution.

In this paper, we proposed an algorithm to determine cutoff threshold to select an optimal feature subset using collaboration with statistical outlier detection. We present two statistic-based feature scores which are insensitive to the imbalanced class distribution based on normalization of the Document Frequency and evaluation of the difference between two sample means. We also introduced the method to reduce the effect of imbalance class dataset by separating the main classes into many subclasses and use them instead of the main class for computing feature score. Moreover, we proposed the algorithm to evaluate and eliminate redundant feature based on measuring correlation between features. Compared to feature sets selected by existing methods, the experimental results show that performance of a feature set selected by the proposed methods are comparably equal and better when it is tested on the standard dataset such as Reuter21578, RCV1v2, etc. and the optimal feature set is a compact size.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature