Feature selection for high dimensional data: An evolutionary filter approach.

Abstract

Problem statement: Feature selection is a task of crucial importance for the application of machine learning in various domains. In addition, the recent increase of data dimensionality poses a severe challenge to many existing feature selection approaches with respect to efficiency and effectiveness. As an example, genetic algorithm is an effective search algorithm that lends itself directly to feature selection; however this direct application is hindered by the recent increase of data dimensionality. Therefore adapting genetic algorithm to cope with the high dimensionality of the data becomes increasingly appealing. Approach: In this study, we proposed an adapted version of genetic algorithm that can be applied for feature selection in high dimensional data. The proposed approach is based essentially on a variable length representation scheme and a set of modified and proposed genetic operators. To assess the effectiveness of the proposed approach, we applied it for cues phrase selection and compared its performance with a number of ranking approaches which are always applied for this task. Results and Conclusion: The results provide experimental evidences on the effectiveness of the proposed approach for feature selection in high dimensional data.

Keyword: Genetic algorithm; Feature selection; High dimensional data; Filter approach; Machine; Learning (ML); Evaluation function; Proposed approach; Search algorithm; Natural; Language processing.