ABSTRACT

The main focus of this thesis is on Knowledge Discovery in Databases (KDD). This thesis presents novel solutions for enhancing the performance of classification algorithms. The proposed solutions yield significant improvement in the performance of classification by adapting hybrid feature selection approach. The performance of a classification algorithm in data mining is greatly affected by the noisy information such as redundant and irrelevant features. These parameters of data not only increase the cost of mining process but also degrade the quality of the result. Five different research works related to KDD namely, Study of Ranking methods, Feature Selection using Genetic Algorithm, Hybridization of Symmetrical Uncertainty ranking method with Genetic Algorithm , Hybrid Feature Selection using ReliefF ranking with Entropy based Genetic Algorithm for breast cancer prediction and Symmetrical Uncertainty ranking method with Particle Swam Optimization for heath care analytics have been investigated.

Feature selection is a fundamental problem in data mining to select relevant features and cast away irrelevant and redundant features based on some evaluation criteria. A subset selected by the feature selection algorithm would have better predictive accuracy than the model built with a complete set of features. Selecting the right set of features for classification is one of the most important problems in Data mining. In this thesis, improved data classifications by hybrid feature selection through the ranking with evolutionary algorithms have been proposed. ReliefF and Symmetrical Uncertainty ranking methods effectively remove redundant features and then the evolutionary algorithms are applied to select optimal relevant features in order to maximize the classification accuracy. Next, the reduced dataset is used to build classification models using well known classifiers. The classification accuracy has been evaluated by a set of data mining metrics. The influence of the proposed feature selection methods are evaluated by well known classifiers. The experiments on complex medical

are evaluated by well known classifiers. The experiments on complex medical and breast cancer data demonstrate that the proposed methods are robust and effective approaches which can find subsets of features with higher classification accuracy and/or smaller size compared to each individual feature selection algorithm.

Overall, this thesis deals with the performance of different classification algorithms and the impact of hybrid feature selection method on these popular classifiers. This proposed hybrid algorithms are much faster and scale well the data sets in terms of selected features, classification accuracy and running time than most of the existing algorithms.