Abstract

Classification in data mining assigns items in a collection to target classes. The goal of classification is to accurately predict the target class for each case in the data. A classification task begins with a data set in which the class assignments are known. In classification models, a classification algorithm finds relationships between the values of the predictors and the values of the target. In this thesis, three modified classification techniques have been used for more accurate anomaly detection leading to better detect diabetes, cancer, credit card fraud and network attacks. These techniques are 1. Probabilistic Feature Engineering 2. Classification Accuracy Enhancement through Meta-learners 3. Re-engineering of ensembles for real world datasets. In the first technique, statistical control chart combined with modified decision tree algorithm has been used. The data sets were initially cleaned for missing values. We implemented a statistical control chart approach and modified decision tree algorithm to solve anomaly detection problem. Proposed algorithm uses modified entropy measure for selecting relevant attributes from the data sets. Here, before applying the modified entropy measure, each attribute is tested against the statistical control bound limits. In the second technique, an optimized probabilistic based feature selection model was implemented on credit card data set for fraud detection and network dataset for detection of malicious attacks. This model efficiently detects the anomaly features along with uncertain features with high true positive rate. Experimental results show that proposed approach outperforms well against traditional anomaly optimization techniques and it performs well against different distributed datasets in terms of time and accuracy. In the third technique, Ensemble modeling is used. This model is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics. A new feature selection based ensemble model was proposed on high dimensional microarray disease datasets. Most of the conventional classification techniques deal with limited attributes and small datasets. Proposed model is one of the ensemble learning models, which is capable to handle datasets with large number of attributes. In conclusion, our modified classification techniques applied on large data sets and distributed data sets have yielded fruitful results for better

detection of anomalies leading to more accurate detection of diabetes, cancer, fraud detection in credit cards and malicious attacks in network communication.