A Novel Method For Reducing Risk With Optimized Anomaly Classifier

C. Kavitha
Dept of Computer Science
Madurai Kamaraj University College,
Madurai, Tamil Nadu, India
kkavitha009@gmail.com

R. Jayakarthick
Dept.of . Computer Science
School of computing Sciences
Vels institute of science technology and advanced studies,
Chennai, Tamil Nadu, India.
drrjayakarthik@gmail.com

M.S. Nidhya
Dept of Software engineering
PeriyarManiammai Institute of Science & Technology,
Thanjavur, TamilNadu, India.
nidhyaphd@gmail.com

Abstract -Deciding on an uncertain event may lead to risk. Uncertainty occurs due to the lack of knowledge of a particular event or a situation. The only way to avoid this is to analyze the Risk. The risk analyzed properly will reduce the impacts to a great extent. Here the risk management dealt with the anomaly detection mechanism, which is carried out by clustering the data to find the outlier as the anomaly.The proposed method employs the PCA mechanism for dimensionality reduction which is further clustered with K means algorithm and classified with decision tree algorithm.

Key words: PCA, clustering, classifier, anomaly.

I INTRODUCTION

Risk identification is the process of listing potential risk and their characteristics which is more important for risk analysis. It is usually done at the beginning stage of the project. An individual or company takes the risk only if it will help to achieve its target while keeping all under control, which is considered to be a risk. To avoid the risk first, we have to analyze the potential issues that could bring negative impact, which would help the organization to avoid risk.[1] Explained how enterprises have to face an international market by identifying all kinds of risk based on identification, classification, and empirical measures of the risk content. Risk management can be successful only with proper planning and execution. Risk identification, Risk Mitigation, and Risk application are analyzed for transportation projects[2].Here the risk reduction is carried out with data mining techniques. Data mining is the collection of exploration techniques based on advanced analytical methods and tools for handling a large amount of information.It is used to discover pattern from the current data and compare it with earlier set of data to find the change.A [3] summarized survey on datamining was undergone and the way to achieve neural network and genetic algorithm using the data mining technique was also studied.

II PREPROCESSING

Preprocessing is the method of processing the original raw data for further proceedings. The original data is inconsistent, incomplete, and contains noise. The data extracted from the larger data set is cleaned and filtered according to their need. [4] used feature extraction for preprocessing the email and its URL from the dataset and used Naïve Bayes classifier for detecting the spam mail. Here the preprocessed data will be more precise when used with a classifier which improves the prediction accuracy than used with the non-preprocessed dataset.[5] evaluated the data preprocessing on blog articles in the Slovak language where lemmatization is used to enhance the quality of clusters.

III BACKGROUND STUDY

3.1. Anomaly

One of the important problems in the research is the identification of abnormal instances. Identifying these abnormal instances play an important role in the application of risk reduction, fraud detection, and network security. [6] used a two-pass technique to find the anomaly. The first pass is used to find clusters with k means algorithm and the second pass uses the ACO technique to refine the clusters and to find outlier as the anomaly. [7] used k means for analyzing and visualizing the flow of data to detect anomalies in the network using the data attribute as IP address, protocols, and port number.SathyaNarayana et al. [8] used unsupervised and domain-independent such as UNICORN to utilize the information provided by various links to find the anomaly.

3.2. Principal Component Analysis

It reduces multidimensional data into lower dimensional data. For high-dimensional datasets, dimension reduction is usually performed before applying a clustering algorithm to avoid the effects of the occurred due to dimensionality. PCA is one of the feature extraction methods to capture the underlying variance from the large dataset with orthogonal linear projection.Random data selected from the original dataset to prepare a transformation matrix and shifted