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

2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0521-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICACCS51430.2021.9441799

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

978-1-6654-0521-8/21/$31.00 ©2021 IEEE

the transformation matrix of reduced dimensions. This will project the original data into the new dimension because with large dimensional data it is difficult to estimate accuracy [9, 10].[11] used the PCA method for dimension reduction on a fast-food dataset to predict the diseased and used k means algorithm for grouping their data.

3.3. K means clustering

It is the process of a grouping of similar objects into the same class. K means to come under the unsupervised learning algorithm. In k means clustering algorithm first the number of clusters is specified. Then the initial centroid is randomly selected from a large dataset. The squared distance between the centers is calculated and the data is assigned to the closest centroid. In [12] used an optimized framework for detecting the network attack using k means and decision tree algorithm. They proved that the detection rate is higher than the false alarm rate. [13] proposed k means algorithm to analyzed social media with optimized cluster distance along with genetic algorithm and showed the benchmark result on WSS comparison with the already existing one. [14] used k means algorithm along with PCA based genetic classifier which showed that the proposed method gives a significant result with the performance metrics. [15] used the ICV technique to find inter-cluster similarity and dissimilarity. They used this technique in four unsupervised clustering algorithm to find the inter-cluster similarity and dissimilarity. The results showed that the ICV technique is more significant in finding the inter clusters simply.

3.4. Classifier

The technique to divide the data into several classes is called classification. The classification can be done on structured or unstructured data. An algorithm that matches the given data in a dataset to a particular category is called the classifier. The ability of the classifier depends on its accuracy. Yaguangwang et al. [16] compared four text Classifier based on their speeds and efficiency. They also showed that the Naïve Bayes classifier has a higher accuracy rate than the other three on movie reviews in NTLK. Developed a hybrid algorithm [17] by combining Naïve Bayes and Decision tree algorithm. Used the Decision tree algorithm to divide the dataset based on nationality and applied the Naïve Bayes algorithm on its leaf node. The classification accuracy of the hybrid classifier is higher than the other two individual results. [18]compared five types of flavivirus to study about zika virus which comes under this flavivirus genus. They used Apriori and K means algorithm to group the common character of the flavivirus and Decision Tree Algorithm to identify the odd one. The Decision tree classifier is used in the proposed system. [14] used PCA for dimensionality reduction and two pass clustering refined by ACO. Here they used the Decision tree algorithm as classifier guided by Genetic algorithm.

Decision Tree Classifier

It is one of the simplest methods to classify the given data. It comes under a supervised machine learning algorithm. By choosing the root with the highest information gain and keeping all other values in its branches, the decision tree algorithm works.The internal node is chosen recursively and the process continues. The tree gets stopped when there are no more nodes for separation and when all samples belong to the same class.

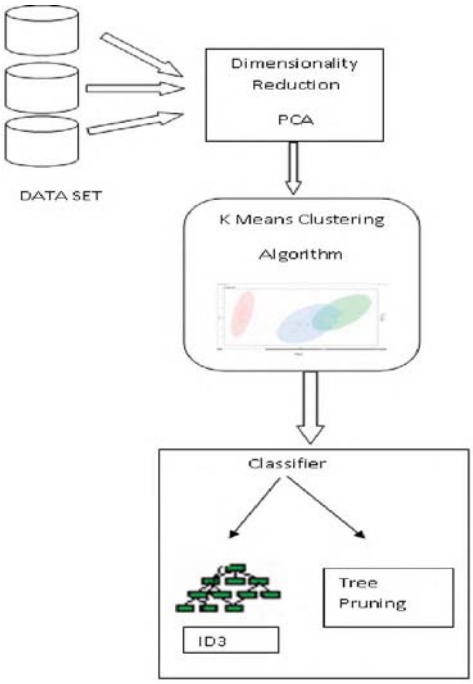

Fig. 1. ARCHITECTURAL DIAGRAM FOR THE PROPOSED SYSTEM

IV PROPOSED APPROACH

As follows, the algorithm for the proposed approach is given.

INPUT: Breast Cancer dataset

Step 1: Preprocessing is done by normalisation of the Min-Max

/*formula

$$z = \frac{n - \min(n)}{[\max(n) - \min(n)]}$$

*/

Step 2: Application of PCA for the reduction of dimensionality

/* PCA algorithm

1. Mean data centre
2. Find the Dimension Covariance Matrix
3. Find proprietary vectors for a covariance matrix
4. Decreasing values of Eigenvalues are considered to sort eigenvectors
5. The highest Eigenvalue of the eigenvector is the Principle component of the data

*/

Step 3:K means algorithm

/*

1. Specify the number of clusters.

2. centroid k is selected randomly from the shuffled dataset.
3. Repeat the process until there is no change in the centroid.
   - The squared distance between centers and the data point is computed.
   - Each data point with minimum distance is assigned to the closest center.
   - The center for the cluster is computed by finding the average of the data point that corresponds to each cluster. */

Step 4: Clustering-based Multiclass classifier

/* ID3 algorithm to build the decision tree

(a) Tree construction

The algorithm for decision tree is given as follows

1. Select an attribute as the root with the highest data gain and branch all its values.
2. Select the internal nodes (attributes) with their exact values as branches for each cluster.
3. Stop when all specimens belong to the same class, then the tip becomes the leaves of that class or no more specimens remain.

(b) Tree pruning

Identify and separate branches that returnnoise or outliers by K means Clustering

Output: Multi-Class classified dataset

V RESULTS AND DISCUSSION

The Breast cancer dataset is used for analysis which consists of 32 attributes. The first column consists of the Id of the patient and the second column covers the details that the patient has a malignant tumor and beginant tumor. Fig.2 shows the normalized data of the given dataset.

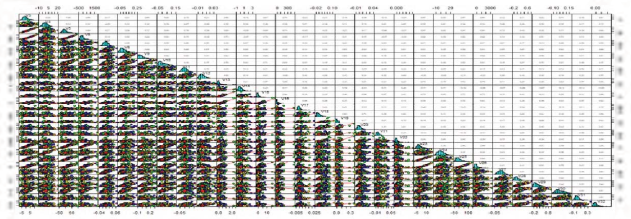

Fig. 2 Normalized data

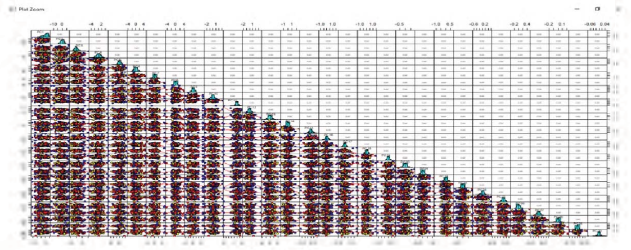

Fig. 3 [illegible]

Fig.3 shows the orthogonality of the Principle Component Analysis where the correlation coefficient is zero which eliminates the problem of multicollinearity.

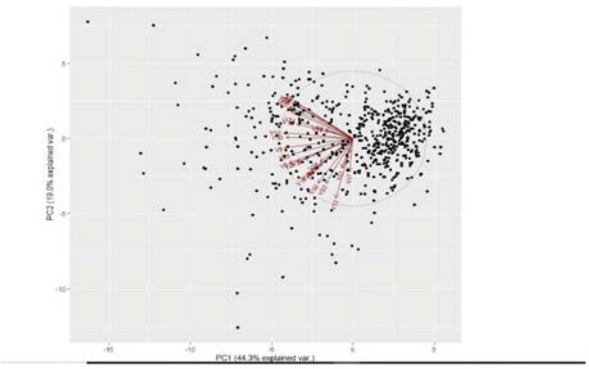

Fig.4 Biplot

Fig.4 is the Biplot of the Principle Component Analysis.

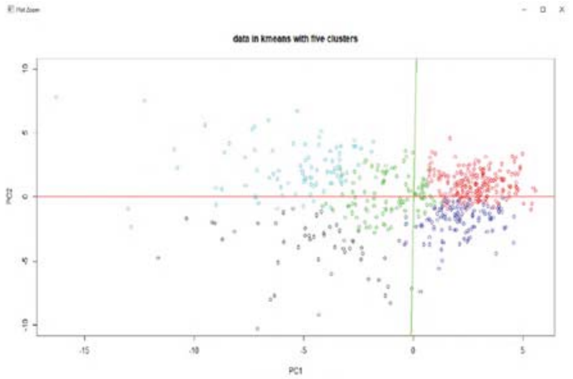

Fig. 5 kmeans with five cluster centre

Fig.5 shows the data plot for k means algorithm with five cluster centers. PC1 which is red covers the highest percentage of the data than the green color which is drawn along the Y- axis represents PC2.

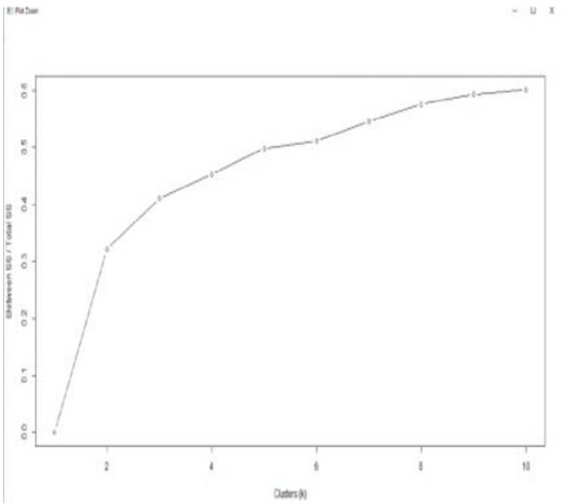

Fig.6 Kmeans with various cluster centres

The Fig.6 shows elbow curve representation for K means algorithm with various cluster centers, which proves that the increase in cluster center increases the sum of the square between the centers.

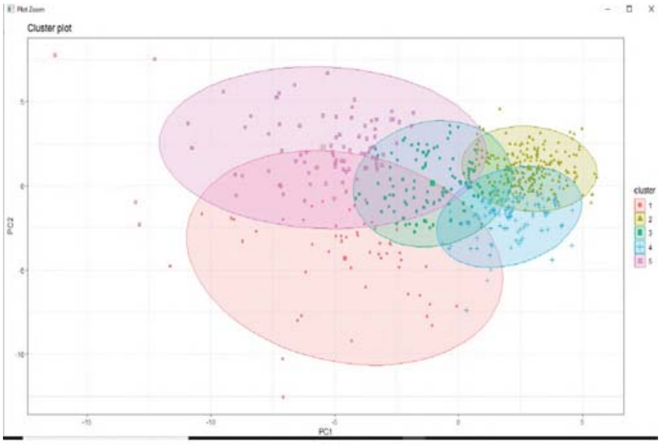

Fig. 7 Cluster Plot

Fig. 7 shows the cluster plot for five cluster centers.

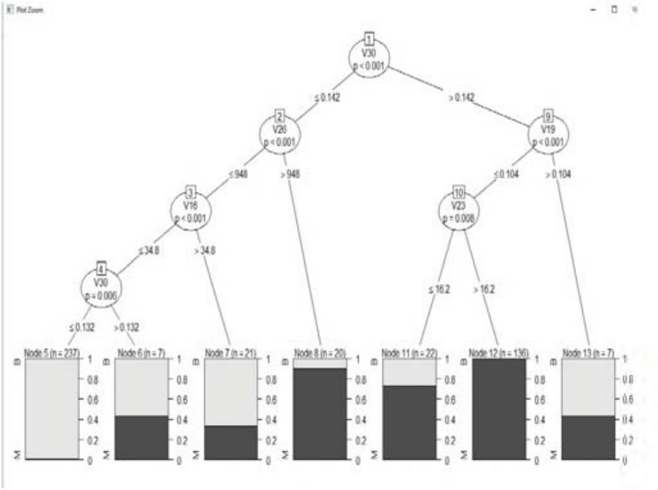

Fig.8 Decision Tree

Fig. 8, Decision Tree drawn with thirteen nodes. The node 5(n=237) represents the condition of Malignant stage. The node 12(n=136) represents the condition of the Beginant stage.

Table 1 Prediction of Training dataset

P1	B	M
B	257	15
M	8	170

The above Table 1 shows the prediction of the training dataset. The data correctly classified for Beginant tumor is 257 and the malignant tumor is 170, where 15 and 8 are misclassified data.

Table 2 Prediction of Test dataset

P2	B	M
B	89	1
M	3	26

Table 2 shows the prediction for test data. The correctly classified data for the Beginant tumor is 89 and the malignant tumor is 26.

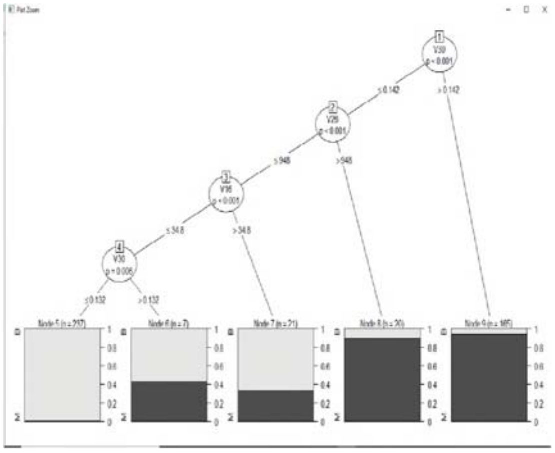

Fig.9 Pruned Decision Tree

Fig.9 shows the Pruned Decision Tree with nine nodes which is smaller, less complicated, and more compact to predict.

Table 3 Prediction of Pruned Training dataset

P1	B	M
B	253	12
M	12	173

The table 3 shows that the misclassified data is reduced after pruning.

Table 4 Prediction of Pruned Test dataset

P2	B	M
B	89	1
M	3	26

The above table 4 shows the correctly classified data for test dataset of pruned tree.

VI CONCLUSION

This paper proposed the risk reduction by employing dimensionality reduction with cluster analysis for anomaly detection. The accuracy of the work is done with the Decision Tree classifier. Thus the risk reduced by the proposed method is more significant. The work can be further done with different classifiers.

Reference

[1]. XU Hui.WAN Yi-qian.: "Risk Identification and Measure Based on Data Analysis —Take Internationalization Risk as an Example", International Conference on Management Science & Engineering, Moscow, Russia,(16th)September14-16, 2009.DOI: https://doi.org/10.1109/ICMSE.2009.5317486.

[2]. Paul Franklin., Arup.: "Risk Analysis for Transportation Projects", Annual Reliability and Maintainability Symposium, 26-29, Jan. 2009. DOI:https://doi.org/10.1109/RAMS.2009.4914688.

[3]. Nikita Jain., Vishal Srivastava., "DataMiningTechniques.: A Survey paper", International Journal of Research in Engineering and Technology, Volume 02, Issue 11,pp. 116-119, Nov-2013.

[4]. Haldorai, A. Ramu, and S. Murugan, "Social Aware Cognitive Radio Networks," Social Network Analytics for Contemporary Business Organizations, pp. 188–202. doi:10.4018/978-1-5225-5097-6.ch010

[5]. R. Arulmurugan and H. Anandakumar, "Region-based seed point cell segmentation and detection for biomedical image analysis," International Journal of Biomedical Engineering and Technology, vol. 27, no. 4, p. 273, 2018.

[6]. Kavitha,C., Iyakutti,K.: "Qualitative Risk Analysis through Anomaly Detection by Two Pass Clustering Technique", International Journal of Computational Intelligence Research, ISSN 0973-1873 Volume 9, Number 1, pp. 19-30, (2013).

[7]. Mayank Pal Singh., Subramanian,N.: "Visualization of Flow Data Based on Clustering Technique for Identifying Network Anomalies", IEEE Symposium on Industrial Electronics and Applications, Kuala Lumpur, Malaysia, pp. 973-978, October 4-6, 2009.

[8]. SathyaNarayana., Prasad., Srividhya., and PanduRanga Reddy.: "Data Mining Machine Learning Techniques – A Study on Abnormal Anomaly Detection System",International Journal of Computer Science and Telecommunications, Volume 2, Issue 6, pp.8-14, September 2011.

[9]. VidyaBanu,R.,Nagaveni, N.: "Preservation of Data Privacy using PCA based Transformation", International Conference on Advances in Recent Technologies in Communication and Computing, pp.439-444, 2009. DOI 10.1109/ARTCom.2009.159

[10]. HanumanthaRao,K., Srinivas,G., AnkamDamodhar, VikasKrishna,M.: "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms", International Journal of Computer Science and Telecommunications, Volume 2, Issue 3, June 2011.

[11]. Mohanapriya., Lekha.,Thilak., Mohamed Meeran.: "A novel method for culminating the consumption of fast food using PCA Reduction and K-means Clustering Algorithm",International Conference on Intelligent Sustainable Systems (ICISS), pp.549-552,21-22 Feb. 2019.

[12]. ElhamAriafar., RasoulKiani.:" Intrusion Detection System Using an Optimized Framework Based on Datamining Techniques", IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, Iran, pp. 0785-0791, Dec2017.

[13]. Ahmed Alsayat., Hoda El-Sayed.: "Social Media Analysis using Optimized K-Means Clustering", IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), USA, 8-10 June 2016.

[14]. kavitha,C., Iyakutti,K.: "Multiclass classifier using clustering guided by PCA for Anamoly based Risk reduction", International journal of Applied Engineering and Research, ISSN 0973-4562 Volume 10, Number 4, pp. 8887-8901, (2015).

[15]. Krishnamoorthy,R., SreedharKumar,S.: "A New Inter Cluster Validation Method for Unsupervised Clustering Techniques", 1st ICCCV – 13, Coimbatore, India, December 20 - 21, 2013.

[16]. Yaguangwang., Wenlong Fu., Aina Sui., Yuqing Ding.:"Comparison of four text classifiers on movie reviews", 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence, pp. 495-498, 2015.

[17]. Lt Col AbhishekLal., C.R.S Kumar.: "Hybrid Classifier for Increasing Accuracy of Fitness Data Set", 2nd International Conference for Convergence in Technology (I2CT), pp.1246-1249,2017.

[18]. Chan woo Kim., Se Hwan Ahn, Taeseon Yoon.: "Comparison of Flavivirus Using Datamining-Apriori, K-means, and Decision TreeAlgorithm", February 19-22, pp.454-457, 2017.