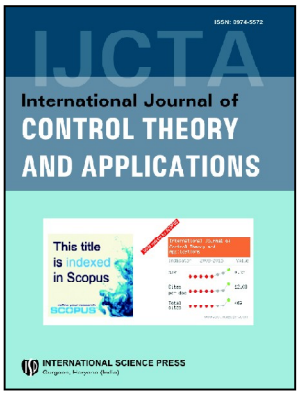

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 29 • 2017

A Novel Filter Based Ensemble Based Anomaly Detection Model for Uncertain Data

[1*]Y. A. Siva Prasad and [2]G. Ramakrishna

[1]Research Scholar, Department of Computer Science and Engineering, KL University, Andhra Pradesh, India*
[2] Professor, Department of Computer Science and Engineering, KL University, Andhra Pradesh, India

Abstract: Due to the rapid growth of high speed network, the risk of credit-card attacks on the complex networks are also increases accordingly. Anomaly discovery from the database is a process of filtering uncertain features, so that it can be used wide variety of applications. Since the online distributed data is the communication between the remote client and the centralized server, it is difficult to predict the occurrence of an anomaly in the large distributed data. Anomaly detection on the complex data must take a long time due to the large number of features. A large number of anomaly prediction models have been implemented in the literature to find the anomaly patterns or features using association and classification techniques. Unfortunately some anomaly detection models over data mining cannot cover all the normal or abnormal features. Traditional approaches mainly focus on detecting relevant patterns from the trained data in order to estimate the test data instances. In this proposed work, the robust partition based classifier is implemented to find the topmost anomalies using attribute relationships. This model efficiently detects the anomaly features along with uncertain features with high true positive rate. Experimental results show that proposed approach has high computational efficiency and anomaly detection rate compared to traditional anomaly detection techniques.

Keywords: Anomaly Detection, Uncertain Data prediction, credit-card attacks, data classification, and filter based classification.

1. INTRODUCTION

For many online applications, machine learning models are required to predict anomalies (abnormal, unexpected or unmodeled) data or features [1]. The success of machine learning model for a particular task is significantly affected by the quality of given complex data. Any noisy, inappropriate, superfluous data may lead to unpredictable results [2]. Real time data generated through sensors, instruments, online applications or measuring errors due to failure while delivering the results [3]. In order to predict the data from such anomalies, different models of data filtering and prediction techniques have been used to lower data complexity and to optimize the accuracy rate [4].

The process of machine learning is an integrated part of developing effective prediction models in the complex applications. However, the main issue with these models is extended data filtering and preprocessing due to redundancies, inconsistencies at the source node to the centralized system in the distributed environment

[5]. Network Intrusion detection is the method of detecting anomaly events occurring in computer system to predict network vulnerabilities and other network issues [6]. Intrusion detection system can be categorized according to the data type and size of the networks. The first type of attack is directly monitored in the host based mechanism in order to find host resources of a particular attack. Whereas, in the network based systems monitor network data using a series of sensors associated to the network in order to detect any malicious activities [7, 8]. Network problems can affect various security requirements, including authorization, authentication, availability and integrity. Attackers can cause various types of network attacks such as worms, viruses, compromises, scan and Denial of Services (DoS) [9, 10].

Anomaly detection may refer to an unsupervised model that produces a data mining model for identifying instances that deviate from the normal in a data. An anomaly detection model is a one-class classification which is used to describe the feature relationship in the distributed data. The basic anomaly detection models on the whole follow the following steps [11, 12]:

- Identify the number of feature vectors to classify the instances.
- Determine the metric to compute the degree of deviation from the instance set.
- Set some threshold measure which exceeds the metric computation is considered as anomaly.

The application of k-means clustering model along with the outlier detection technique has very low true positive rate. Later, k-means with PCA model was implemented to find the feature based anomaly detection [13].

As the volume of data and features increases, traditional data mining models fail to detect the anomalies for boosting the accuracy and efficiency [14].

Problems in anomaly detection models

The key challenges of traditional anomaly detection models on various applications include: medical, credit card, stock market and other complex realtime applications include [15, 16]:

1. Type of the anomaly: It indicates the variation in a value or context anomaly when a value is normal or abnormal.
2. Data type: Data can be univariate or multivariate according to the number of features and its types.
3. Training data: Type of input training data and its size.

This paper is organized as follows: In section II, we will give summarized information regarding the traditional anomaly detection techniques in various applications. After this in Section III, we will discuss the proposed anomaly detection model on the complex data. Section IV presents the experimental results.

2. RELATED WORK

Network anomaly detection using data mining approaches is used as detection techniques against network threats. Different types of data used for network intrusion detection such as logs, system calls, traffic data, network features, etc. Most of the traditional anomaly techniques use audit data, log files and packet data to predict network attacks. Unfortunately, this most of the traditional approaches fail to predict the attack using log files and audit data. Techniques like Decision tree, Naïve bayes, neural network and SVM are used to classify the network instances based on training data. Even though each technique has its own strengths in finding network attacks, there are also limitations in several detection techniques, usually takes time to build the model, and takes time to load the large volume of data and high false positive rate. Different types of classification mechanisms have been implemented in [17] with a series of multi stage classifier to overcome the limitation on high false positive rate. In each stage, binary classifier is used to reduce the network features by partitioning the data into

single class and normal type of attacks. This system can reduce the false rate but fails to predict the new type of network patterns. Finally, the Multi-layered network is compared to single layered network for network attack detection. Multiple Hidden Markov models are implemented in the literature [10] to identify anomalies in the series of credit card records [18]. Table 1, summarized the credit card detection models.

Table 1
Credit card anomaly detection models

Main Objective	*Algorithm*	*Handle Mixed Attributes*
(C) Comparison of three classification models by using history data. [Data set: not identity][6]	Decision tree, Neural Networks, Logistic regression	No
(T) Analyze and filter customer's behavior for real-time fraud detection. [9]	Self-organizing map (SOM) combined Gaussian function	No
(T) Credit card fraud detection was used by Hidden Markov Model.[10]	Hidden Markov Model	No
(T) Behavior of withdraw money from bank using Fuzzy logic for distribute data and Self-organizing map for detection algorithm[14]	Fuzzy logic combine Self-organizing map	No
(T) Data mining was the fuzzy association rule that extracted behavior patterns may be obtained in fraud transaction.[17]	Fuzzy Association rule	No
(T) Develop an application of genetic algorithm for fraud detection.[18]	Genetic algorithm	Yes
(C) Compare performance was 7 classification models of fraud[20]	CART, Decision tree, chi-squared automatic	Yes

Anomalous series detection and contextual abnormal subsequence detection. In the current research, only real-valued time series are actually categorical and time series are out of the coverage. The main gap between these two techniques is: the first one works to find out which series is anomalous as the latter one wants to know when abnormal behaviour. The problems occur in time series data are to summarize the useful historical details is a difficult problem. The behaviour of outliers is different for different applications, and it makes detecting abnormal behaviour a hard activity. Within a single application domain, the outlier is likewise changing with time hence it needs an effective technique for dynamic prediction [17]. In [6] proposed the DB(pt, dt) Outlier detection scheme, wherein an object obj is said to be an outlier if the distance of pt greater than dt of an object obj. They defined several techniques to find such objects. In this way objects ought to be in relation to those from neighbouring cells to examine if they're outliers [19]. To detect statistical outliers, the residuals of the observations are computed depending on a trained dataset [20]. PCA is a feature extraction and feature selection approach, its mail aim find the relevant feature from the large search space. Main drawbacks identified with the PCA model is: Linear relationships between the large number of credit card features and it doesn't handle dynamic data sets.

3. PROPOSED METHODOLOGY FOR UNCERTAIN DATA CLASSIFICATION BASED ON FILTERING

Uncertain data prediction is one of the complex tasks in data mining. Hence in this paper a novel method of the uncertain data prediction is proposed. In the proposed method a filter based ensemble based anomaly detection is applied. The proposed method contains three stages they are; Data pre-processing, Feature selection process and Classification model. In data pre-processing the missing or inconsistent values from the database is replace. Then in the feature selection phase a novel approach is applied to select the features based on ranking. The last

phase is ensemble based anomaly detection for better classification. The architecture for the proposed method is given in fig 1.

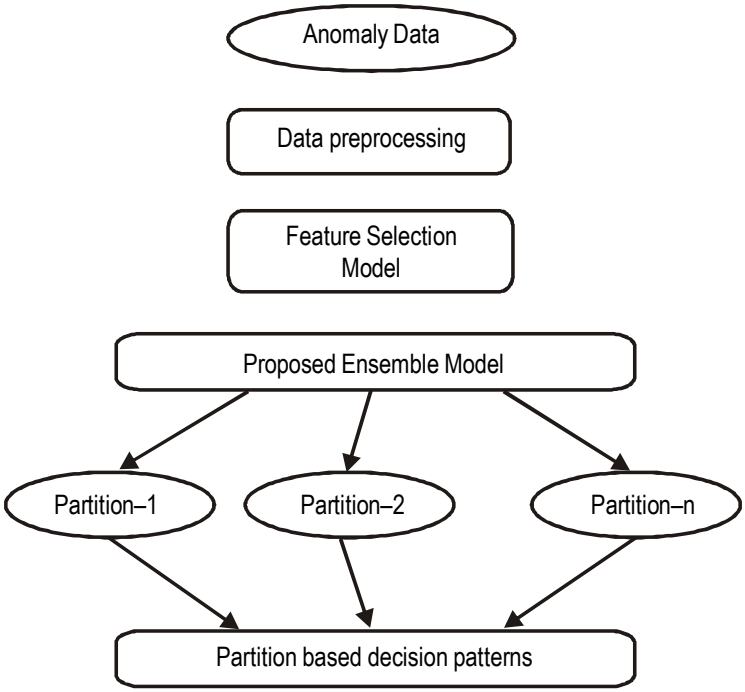

Figure 1: Architecture for proposed uncertain data prediction

(A) Data Pre-Processing

Preprocessing of data is essential in most of the data mining techniques, the process includes in preprocessing are transformation of raw data into an understandable format. An algorithm is applied in Algorithm 1 in this paper for performing data preprocessing. In large data base some data may not be exist and likewise some data might not be predictable. Hence for the better prediction technique it is essential to make understandable, to provide better computation performance.

In the algorithm 1, missing values or inconsistent values are replaced with the computed value. If the attribute is numerical then all the missing values are replaced with the computed Max-Min value. If the attribute is categorical, then all the missing values are replaced with probabilistic ranked value.

(B) Feature Selection

Selection of feature is another important phase in data prediction techniques, in the proposed method a fraud detection attribute selection algorithm is applied for the feature selection. In the proposed algorithm the feature is nothing but the rank. That is a rank value will be calculated from the preprocessed data. The pseudo code for feature selection is given in Algorithm 2.

In the algorithm 2, rank based fraud detection attributes are selected using a novel approach. In this model, entropy and mutual information measures are computed to each attribute with the remaining attributes. Also, similarity measure is computed to all the data partitions for intra cluster variations. The rank of an attribute is computed using the entropy, mutual information and similarity measure. Feature attributes are selected using the user defined threshold.

Read Database D,
For each data record in D
Do
For each feature F in the record
Do
If (F!=NULL) Then Continue;
Else F_type=check_type(F);

If (F_type==numerical)
Then

$$Miss\_Value = \frac{Max(F)\times\sigma_F^2 - Min(F)\times\mu_F^2}{N(N-1)\times[Max(F)-Min(F)]}$$

Value (F)=Miss_Value;
End if
If (type==Categorical)
Then
Freq[]=frequency (F);
// each category of class attribute.
//Probability of each instance value per class.

$$Prob[] = \sum_{i=1}^{m} Prob(x_j / C_i)$$

i=1,2,3…m classes
j=1,2…n instances
rank=Max{freq[]}/Max{Prob[]};
//Fill the value with the max ranked class value.
End if
Done
Done

Algorithm 1: Pseudo code for Data Preprocessing

Input: Filtered data FDB
Output: Ranked feature attributes

For each filtered feature ff in FDB
Do
Compute entropy E(ff);
Compute mutual information between attributes.
$MI(ff) = Max\{MI\{ff, F-ff\}\};$
//Partition the feature ff into m classes as
//Find the similarity between instances of two distinct partitions as

$$Sim(p_i, p_j) = \frac{2\times\sum_{i,j}|x_i - x_j|^2}{N_i(N_j - 1)}$$

//Where N_i is the number of instances in i[th] partition and N_j is the number of instances in j[th] partition.
//Rank of the attribute is defined as
$R(ff) = E(ff) + MI(ff) + Max\{Sim(p_i, p_j)\}$
Done
Input k as user defined threshold
For each r in R (ff) do
If
(r>k)
Then
(Select as fraud feature attribute)
Else
Continue;
Done

Algorithm 2: Pseudo code for fraud detection attribute selection algorithm

(C) Proposed uncertain data prediction

In the proposed method, an ensemble based anomaly detection model is used for the prediction of uncertain data. The step by step procedure of the proposed model is given as follows.

Step 1: Data Initialization

An uncertain database in initialized in the proposed model, which can be further processed for the further prediction process.

Step 2: Preprocessing

The second stage of our proposed prediction technique is preprocessing, in which the missing data value is replaced. Data transformation is applied in the proposed model to consolidated raw data into appropriate form. The transformation for unequal distribution as follows.

For each attribute A_i in database, if the type of attribute Ai is numerical then the value become as in eqn. (1).

$$A_i.value = \frac{A_i.value + G.M(A_i)}{(Max[A_i.value] - Min[A_i.value])} \times (ScaleMax - ScaleMin) \quad (1)$$

For each randomized sample Si, calculate similarity value based on the eqn. (2).

$$S = \sum_{i,j} Sim(S_i, S_j);\ \forall\ Sim(S_i, S_j) > 0 \quad (2)$$

Where;

$$Sim(S_i, S_j) = \begin{cases} 0 & if\ i = j \\ \frac{\pi}{2}\sum(x_i - x_j)^2 \times e^{-(|x_i - x_j|)^2 / 2\sigma_x} & if\ i \neq j \end{cases}$$

Step 3: Anomaly Detection

In this stage the uncertain data is predicted, the partitioning model is applied in the proposed technique. Initially compute prior probability for all S_i in S, the formula to calculate prior probability is given in eqn. (3).

$$\text{PProb} = \arg\max_{i=1,2..m} \frac{\text{P}(c = c_i) \prod_{j=1,2,..n} \text{P}(S_j(j)/c_i)}{\text{P}(A_i = a_1, a_2, ...a_k)} \tag{3}$$

Divide the partition into 2 classes as yes and no or true or false as in eqn. (4). Let P and N are the two sample instances with positive and negative classes.

$$\left.\begin{aligned} P &= \{x_1, x_2, ...x_{N1}\} \\ N &= \{y_1, y_2, ...y_{N2}\} \end{aligned}\right\} \tag{4}$$

Then the True information entropy is computed as in eqn. (5).

$$\text{E}(\text{P}) = \text{E}(\{x_1, x_2, ...x_{N1}\}) \tag{5}$$

Prior entropy function of true samples is given as in eqn. (6).

$$\left.\begin{aligned} \text{TProb} &= \frac{-\sum \text{P}(x_i/c_m)}{\sum \text{P}(x_i/c_m) + \sum \text{P}(y_i/c_m)} \left\{ \log\left(\frac{\sum \text{P}(x_i/c_m)}{\sum \text{P}(x_i/c_m) + \sum \text{P}(y_i/c_m)} \right) \right\} \\ \text{FProb} &= \frac{-\sum \text{P}(y_i/c_m)}{\sum \text{P}(x_i/c_m) + \sum \text{P}(y_i/c_m)} \left\{ \log\left(\frac{\sum \text{P}(y_i/c_m)}{\sum \text{P}(x_i/c_m) + \sum \text{P}(y_i/c_m)} \right) \right\} \end{aligned}\right. \tag{6}$$

Based on the above two probability values the prediction decision is made as be the below rule in fig 2.

```
if (TProb < FProb)
   Anomaly
else
   Normal
end
```

Figure 2: Pseudo code for final decision

In the above ensemble based anomaly detection model, each attribute is checked against the data distribution. If the attribute is not uniform distributed then it was transformed to uniform format. For each attribute in the uniform distributed dataset, instances are partitioned into set of subpartions based on classes. After that, similarity computation was applied on the subpartitions to find the relevant relational anomaly features. The true probability and false probability measures are used to find the high anomaly patterns in each partition.

4. RESULTS AND DISCUSSION

The proposed method for the uncertain data prediction is tested in German credit dataset using JAVA. The German Credit dataset has been obtained from the UCI Repository of Machine Learning Databases. The dataset includes data, which describes about the credit risks of people as good or bad. The dataset contains 1500 samples and 21 attributes, among them 7 numerical and 13categorical attributes. The results obtained from the testing is given in table 2.

Table 2
Results obtained from proposed method

Measure	*Value*
Elapsed time	11.69s
Number of Iterations	9
F-Measure	0.87
Recall	85%
TP rate	0.97
FP rate	0.03
Classification Accuracy	98%

The results shown in the table 2, is acceptable range that is the proposed classifier has around 98% of accuracy and recall about 85%. In order to show the effectiveness of the proposed system, a comparative analysis is made with the existing classifier proposed in [21]. The results obtained from the comparative analysis is given in table 3.

Table 3
Performance Comparison

Measure	*Technique*	
	Existing [21]	*Proposed*
Elapsed time	12.34s	11.69s
Recall	83%	85%
Classification Accuracy	92%	98%

The comparison results from the table 3, show that the proposed technique has better performance than the existing system. The comparison of time take by varying the data size is given in fig 3.

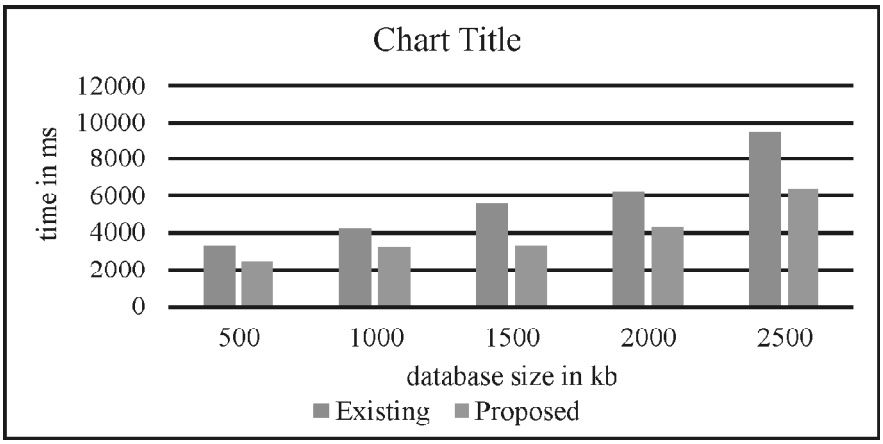

Figure 3: Comparative results for existing and proposed model

In the above fig3, as the database size increases, proposed model has low runtime compare to traditional model. The Comparison between true positive and false negative rate is given in table 4.

Table 3
Comparison between true positive and false negative rate

Database size (kb)	*Tprate*	*Fprate*
500	0.9755	0.042
1000	0.9844	0.0156
1500	0.9854	0.175
2000	0.9785	0.098
2500	0.988	0.154

In the above table, as the database size increases proposed model has high true positive rate compare to traditional ensemble model. Similarly, as the data size increases, proposed model has low false positive rate compared to traditional model. The so far comparative analysis proves that the proposed method has better performance than the conventional method. Thus the proposed data prediction technique will suitable for real time data prediction.

5. CONCLUSION

Several anomaly detection techniques have been implemented in the literature to find the anomaly patterns or features using association and classification techniques. Unfortunately some anomaly detection models over data mining cannot cover all the normal or abnormal features. Traditional approaches mainly focus on detecting relevant patterns from the trained data in order to estimate the test data instances. In this proposed work, the robust partition based classifier was implemented to find the topmost anomalies using attribute relationships. This model efficiently detect 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.

REFERENCE

[1] Geoffrey E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural computation, Vol. 14, no. 8, pp. 1771-1800, 2002.

[2] Shehroz S. Khan, and Jesse Hoey, "Review of fall detection techniques: A data availability perspective," Medical Engineering & Physics, Vol. 39, no. 1, pp. 12-22, 2017.

[3] Jiawei Han, Jian Pei, and Micheline Kamber, "Data mining: concepts and techniques," Elsevier, 2011.

[4] Noha A. Yousri, Mohammed A. Ismail, and Mohamed S. Kamel, "Fuzzy outlier analysis a combined clustering-outlier detection approach," In Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 412-418, 2007.

[5] Francisca Nonyelum Ogwueleka, "Data mining application in credit card fraud detection system," Journal of Engineering Science and Technology, Vol. 6, No. 3, pp. 311-322, 2011.

[6] Ashphak Khan, Tejpal Singh, and Amit Sinhal, "Implement credit card fraudulent detection system using observation probabilistic in hidden markov model," In Proceedings of Nirma University International Conference on Engineering, pp. 1-6. 2012.

[7] Pedro Garcia-Teodoro, J. Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," computers & security, Vol. 28, no. 1, pp. 18-28, 2009.

[8] Tiranuch Anantvalee, and Jie Wu, "A survey on intrusion detection in mobile ad hoc networks," In Wireless Network Security. Springer US, pp. 159-180, 2007.

[9] Addisson Salazar, Gonzalo Safont, Antonio Soriano, and Luis Vergara, "Automatic credit card fraud detection based on non-linear signal processing," In Proceedings of IEEE International Carnahan Conference on Security Technology, pp. 207-212, 2012.

[10] Aihua Shen, Rencheng Tong, and Yaochen Deng, "Application of classification models on credit card fraud detection," In Proceedings of IEEE International Conference on Service Systems and Service Management, pp. 1-4, 2007.

[11] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo, "A geometric framework for unsupervised anomaly detection," In Applications of data mining in computer security. Springer US, pp. 77-101, 2002.

[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey," ACM computing surveys, Vol. 41, no. 3, pp. 15, 2009.

[13] Victoria J. Hodge, and Jim Austin, "A survey of outlier detection methodologies," Artificial intelligence review, Vol. 22, no. 2, pp. 85-126, 2004.

[14] Khyati Chaudhary, Jyoti Yadav, and Bhawna Mallick, "A review of fraud detection techniques: Credit card," International Journal of Computer Applications, Vol. 45, No. 1, pp. 39-44, 2012.

[15] Michał Woźniak, Manuel Graña, and Emilio Corchado, "A survey of multiple classifier systems as hybrid systems," Information Fusion, Vol. 16, pp. 3-17, 2014.

[16] Min Chen, Shiwen Mao, and Yunhao Liu, "Big data: A survey," Mobile Networks and Applications, Vol. 19, no. 2, pp. 171-209, 2014.

[17] Sam Maes, Karl Tuyls, Bram Vanschoenwinkel, and Bernard Manderick, "Credit card fraud detection using Bayesian and neural networks," In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261-270. 2002.

[18] Kenji Yamanishi, and Jun-ichi Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time series data," In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 676-681, 2002.

[19] Hawkins Simon, Hongxing He, Graham Williams, and Rohan Baxter, "Outlier detection using replicator neural networks," In International Conference on Data Warehousing and Knowledge Discovery, pp. 170-180, 2002.

[20] Karim Lekadir, Robert Merrifield, and Guang-Zhong Yang, "Outlier detection and handling for robust 3-D active shape models search," IEEE Transactions on Medical Imaging, Vol. 26, no. 2, pp. 212-222, 2007.

[21] Christopher J. Hutton, Zoran Kapelan, Lydia Vamvakeridou-Lyroudia, and Dragan A. Saviæ, "Dealing with uncertainty in water distribution system models: a framework for real-time modeling and data assimilation," Journal of Water Resources Planning and Management, Vol. 140, No. 2, pp. 169-183, 2012.