

Research Article

Anomalies Calculation and Detection in Fuel Expense through Data Mining

Kaleem Habib and Arif Iqbal Umar

Department of Information Technology, Hazara University, Mansehra, Pakistan

Abstract: In organizations having large vehicle fleet a reasonable portion of the fuel budget is misused by malpractices of drivers and fuel providers. An optimal usage of this fuel amount could be of big advantages to the organization. We proposed a novel anomaly %age calculation algorithm to determine the misuse of the fuel of the vehicles. This algorithm will reduce the efforts for anomaly detection in clustering process. The results reflect that this algorithm could be used to implement an effective check on the misuse of the fuel in big organizations with less effort.

Keywords: Anomaly, data mining, fraud, fuel

INTRODUCTION

Fuel misuse (theft/fraud) is a hot problem in automobile sector. The transport managers do not admit it publicly. They admit privately that at least 15% of the total fuel expenditure is misused. They wish to stop the drain of this huge amount and utilize it optimally to accrue more benefits for the organizations.

Nowadays misuse of vehicles, fuel theft and unauthorized use of vehicle are more attractive activities on part of the corrupt employees and fuel providers because these practices earned them more money due to prevailed high cost of the fuel.

The abnormal trends in the data regarding fuel consummation may reflect suspicious activity (fraud), uneconomical driving or the need for vehicle service.

Data for data analysis could be prepared by employing several different data mining techniques. The Cluster analysis being unsupervised data grouping technique groups similar data in same group on the basis of the distance of the data from the central data item of the group or cluster. Applications of data clustering algorithms depend on situation (data). Different algorithms are effective at different situations.

In this study, we have introduced a new cluster analysis algorithm based on anomaly %age calculation to a dataset provided by a major logistic company in Pakistan. Data set is comprised of the data reflecting Fuel Consumption Values (KMPL- Kilometer per Litter). In our proposed algorithm data points for different vehicle categories are grouped and measure the center point as a mean of all values in each category. These values then transform into their %age of correctness by using KMPL value and center point. The results reflect that the algorithm could be used to implement an effective check on the misuse of the fuel in the big organizations.

MATERIALS AND METHODS

In the accounting literature, most studies focus on management fraud (Thirungsri, 2010). For the prediction of management fraud, most prediction models employ either logistic regression techniques or the Neural Network.

A system is proposed for cluster analysis in Thirungsri (2010), data was cleaned and transformed using SAS. The clean data was then exported into a Comma Separated Value (CSV) file. Then the dataset was prepared in the ARFF format in order to be fed into Weka. New dataset was created based on the original data and Claims with similar characteristics have been grouped together and clusters with small populations have been flagged for further investigations. All this process accumulates high computational and operational cost. Integrity and security of data is also badly lost.

Anomaly detection in the field of Network Security, Information Security, Network Traffic, Node Behavior and wireless Network is mainly depend on Historical data, some pre-defined characteristic and different hardware settings (Shah and Trivedi, 2012; Thottan and Chuanyi, 2003; Ringberg *et al.*, 2007; Janakiraman and Vasudevanf, 2009; Ashoor and Gore, 2011; Pari and Sridharan, 2011; Annie, 2012; Basha, 2012). Classification Methods and Dimensionality Reduction are mainly used for anomaly detection or intrusion detection.

In case of anomaly detection, the unusual behavior or abnormal activities in the network are supposed to be identified (Ashoor and Gore, 2011). Historical data is used to verify the normal behavior of the system. The challenges are; User genuinely changes a system file.

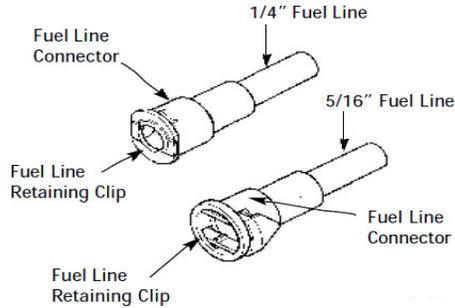


Fig. 1: Visualization of the anti-draw-off devices

Wrong data for the normal behavior can be generated. Sometimes system treats failed logins in a day as abnormal, but some users actually do so normally.

The Department of Detention and Enforcement has conducted an audit to prevent a fuel theft from the East Service Center vehicles (Snelding, 2010). The report observed the following:

- The Fleet Services should install appropriate security devices to prevent or minimize theft of fuel from City owned vehicles. This could include the installation of locking fuel caps and/or anti-draw-off devices in City vehicles.
- The Fleet Services Management should implement policies and procedures to control and limit the use of City-owned equipment.

In order to avoid the vehicle fuel theft, some sensors based approaches are in used, in which five sensors are placed inside and surrounding the petrol tank. If the vehicle is in running condition the petrol flow is very slow, so the sensors ignore it. But if someone stealing the petrol, the petrol flow is very fast. If the petrol flow is very fast then the sensor senses it and sends the message to the Micro controller. Through LCD display one could see the remaining quantity of the petrol in petrol tank. The memory device stores the level of the petrol in petrol tank.

A Network Intrusion Detection System (NIDS) embedded in a Smart Sensor inspired device, under a Service Oriented Architecture (SOA) is proposed in Ramos-Morillo *et al.* (2010). The system is able to operate independently as an anomaly-based NIDS or integrated, transparently, in a Distributed Intrusion Detection System (DIDS).

The sensor based solution is not a cost effective solution in case of the fuel fraud detection and many times these solutions are failed to achieve the objectives.

Another approach is installation of fit-in anti-draw-off devices on the neck of the fuel tank such as NeckIt shown in Fig. 1. It is reported that this approach has resulted in savings of between 2-4% on average on fuel bills. By using this device once the fuel is put in to the fuel tank never draw off from the fuel tank.

In majority cases all of the hardware approaches are failed because these measures can only reduce the opportunity of fuel theft by hard activities with fuel tank. These approaches can't detect fake fuel bill vouchers. Several times it is observed that corrupt employees and drivers bring a fake fuel voucher with fake quantity of fuel as written in fuel voucher. Every time they save some money with the help of pump operators and make a high fuel bill for organization but system can't detect it.

This indicates that all hardware and software based systems are not fully successful to improve the situation. Therefore we proposed a new data mining technique for the solution of the problem. This technique will work as Anomaly %age calculation Algorithm. The propose algorithm will work on a database level and highlight any suspicious entries for further investigation.

The New technique will first calculate KMPL for each point in data set. Then it will calculate the center point of KMPL values for each vehicle category. In next step it will calculate distance of each KMPL value from its corresponding center point. Then anomaly %age will be calculated on the basis of each KMPL value and its distance from corresponding center point. After anomaly %age calculation we can mark any number of values as correct values before clustering process. By this way we need to investigate very less values for anomaly detection with less effort.

New nested-loop Algorithm:

(Pseudo code to detect Anomaly)

Algorithm Starts

1. // variable declaration starts

Create Variable I1 of type number and initialize with 0

Create Variable I2 of type number and initialize with 0

Create Variable I3 of type number and initialize with 0

Create Variable V_OLD_READING of type number

//variable declaration ends

2. Search for All Vehicle Types from dataset

{

Get "Vehicle Type ID" and add to our array of return values

For each Vehicle Type found in the search

Start loop

Increment variable I1 to 1

Set variable I2 value to 0

}

3. Search for all vehicles and possible Fuel Types against each vehicle type

{

Get "Vehicle ID", "Fuel Type ID" and add to our array of return values

For each Vehicle ID found in the search

Start loop

Increment variable I2 to 1

Set variable I3 value to 0

}

4. Calculate KMPL for each record of each vehicle;
5. Identify Center Point "Ci" of KMPL Calculated above for each vehicle
6. Search all KMPL readings for each Vehicle and its Fuel Type
 - {
 - Get "KMPL" and add to our array of return values
 - For each KMPL found in the search
 - Start Loop
 - {
 - Increment variable I3 to 1
 - 7. Measure Distance "Di" of Each KMPL from Center Point "Ci" in step 5.
 - 8. Calculate %age "Pi" of each distance "Di" for correctness of KMPL.
 - 9. Plot a data on the basis of %age Pi calculated from step 8.
 - Repeat steps 7 and 8 for all data points in step 6
 - 3rd loop End
 - }
 - Repeat steps 4, 5 and 6 for all data points in step 3
 - 2nd Loop End
 - }
 - Repeat step 3 for all data in step 2
 - 1st Loop End
 - }
 - Algorithm Ends

RESULTS AND DISCUSSION

This study examined the possibility of calculating the anomaly %age and plot it in clusters for fraud detection in fuel consumption. The anomaly %age calculation algorithm is tested on a data set from a major logistic company in Pakistan. In this algorithm KMPL's value is used to calculate anomaly %age by measuring the distance of each KMPL value from its calculated center point. These tests will be performed on all vehicle categories. This study is a preliminary step to apply anomaly %age value in clusters for fraud detection in automobile industry.

Anomaly calculation: The analysis of the data in Fig. 2, 3, 4, 5, 6, 7 and 8 reveals that, after calculating anomaly %age it is very much clear that it is better way to plot these %ages into clusters for fraud detection and then examined these clusters for further investigation.

Anomaly detection: In this step we will further analyze this data to pinpoint the anomalies and the area where the anomalies exist.

Figure 9, we can see clusters made on the basis of full data set. A lot of anomalies can be seen in this data. Normal clusters are the well-populated area in the figure. Many data elements are below the well-

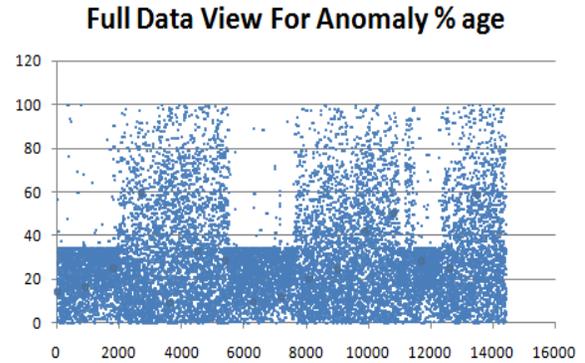


Fig. 2: Visualization of full data with anomaly % age

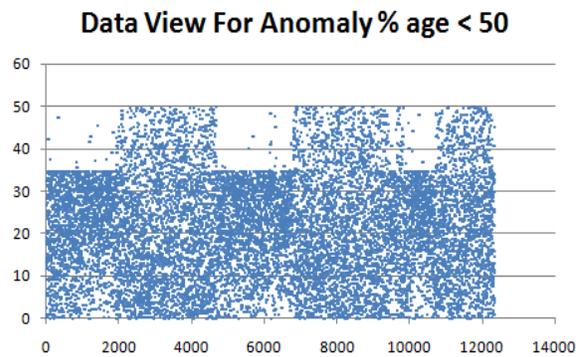


Fig. 3: Visualization of data with anomaly % age < 50

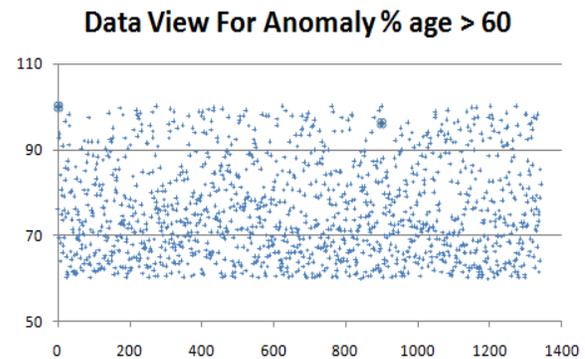


Fig. 4: Visualization of data wit anomaly % age > 60

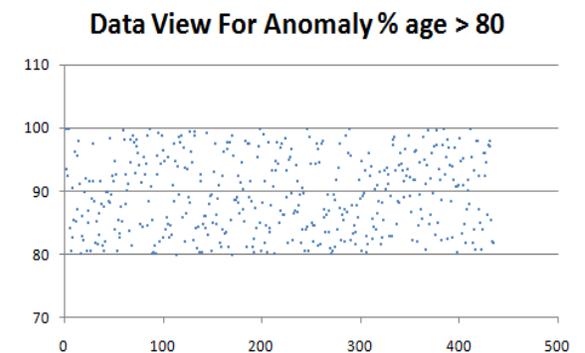


Fig. 5: Visualization of data wit anomaly % age > 80

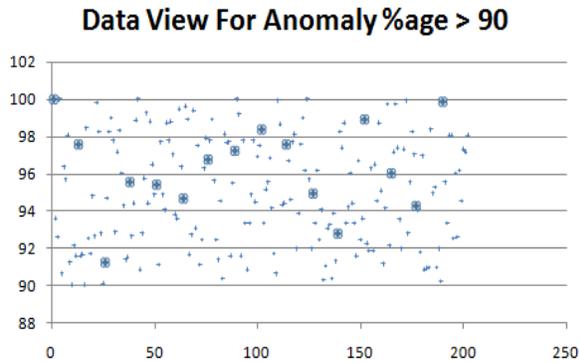


Fig. 6: Visualization of data wit anomaly %age > 90

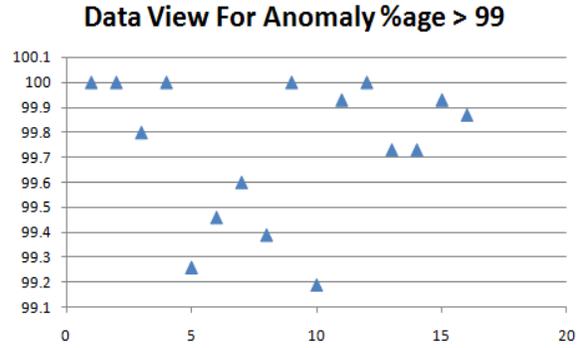


Fig. 8: Visualization of data with anomaly %age > 99

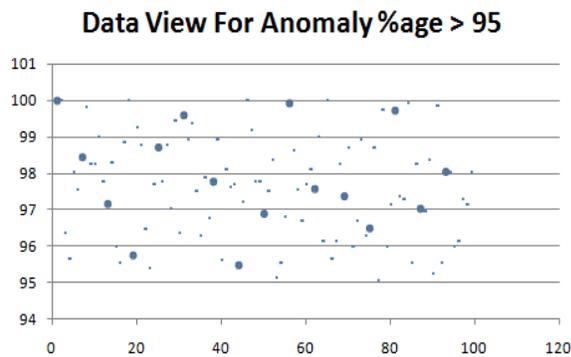


Fig. 7: Visualization of data wit anomaly %age > 95

populated clusters area. These entries can be marked as anomalies for further investigation.

In Fig. 10, we plot a data into clusters on the basis of Vehicle ID's. Most of the vehicles have the normal data set and lies in the well-populated cluster area but a few numbers of vehicles have an abnormal behavior. Vehicle ID's O-90, T-53 and B-36 have a large numbers of anomalies seen in this figure. Some other vehicles have anomalies but that are neglect able.

In Fig. 11, we have plot a data into clusters on the basis of Filling Station. Most of the filling stations have the data in well-populated clusters. Big anomalies are lies in fleet fuel cards data. Another filling station name 3-FS also have a large abnormal behavior in date set. These two stations are marked for anomalies.

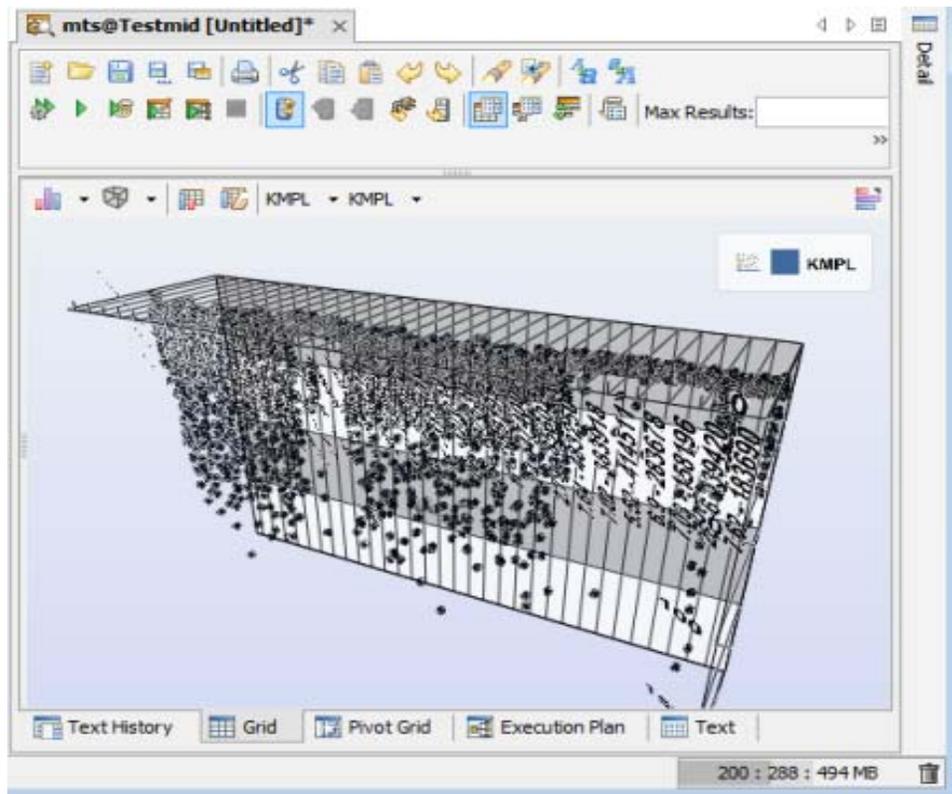


Fig. 9: Cluster analysis of full data with anomaly % age

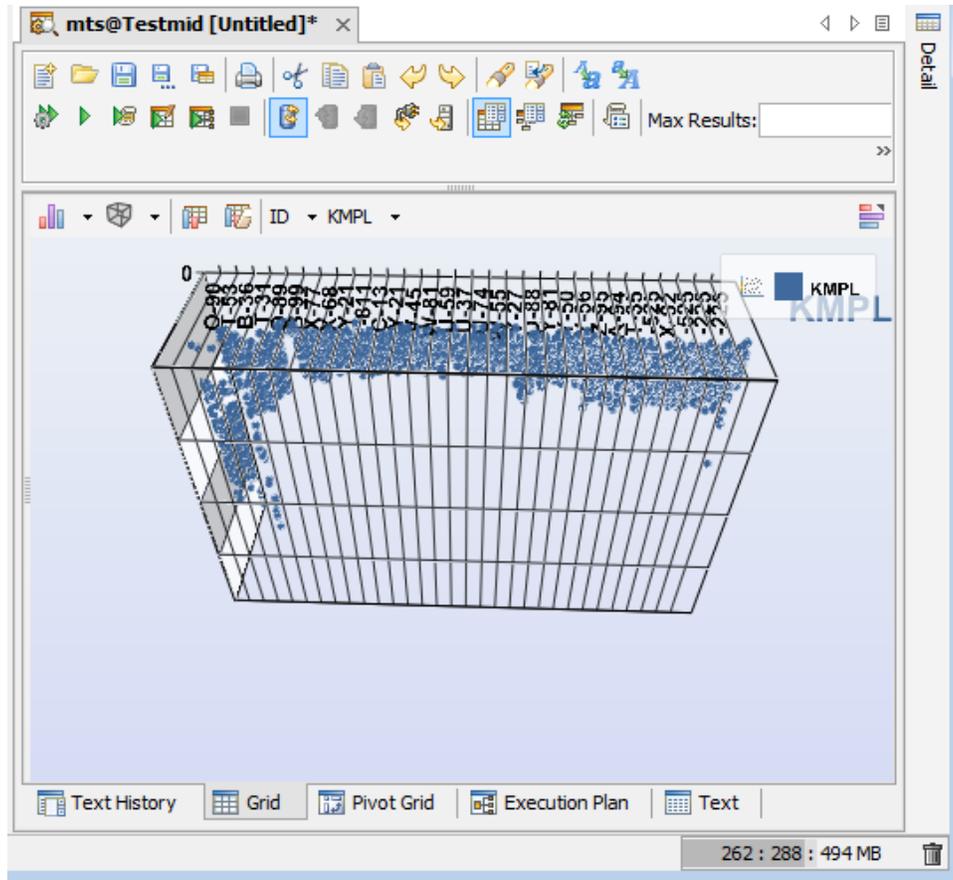


Fig. 10: Cluster analysis on the basis of vehicle ID

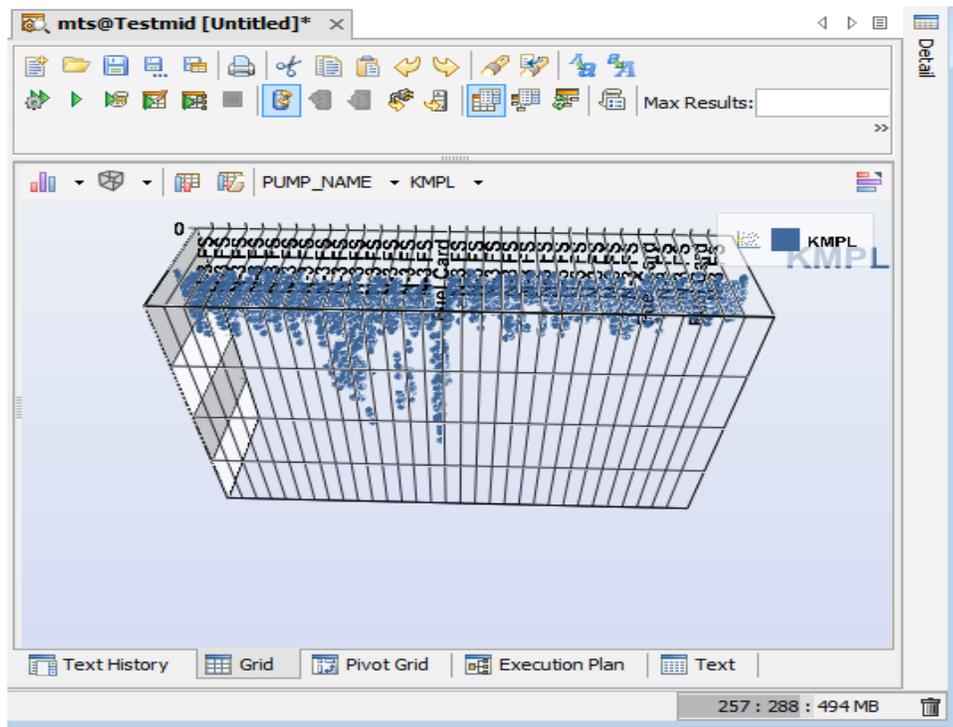


Fig. 11: Cluster analysis on the basis of filling station

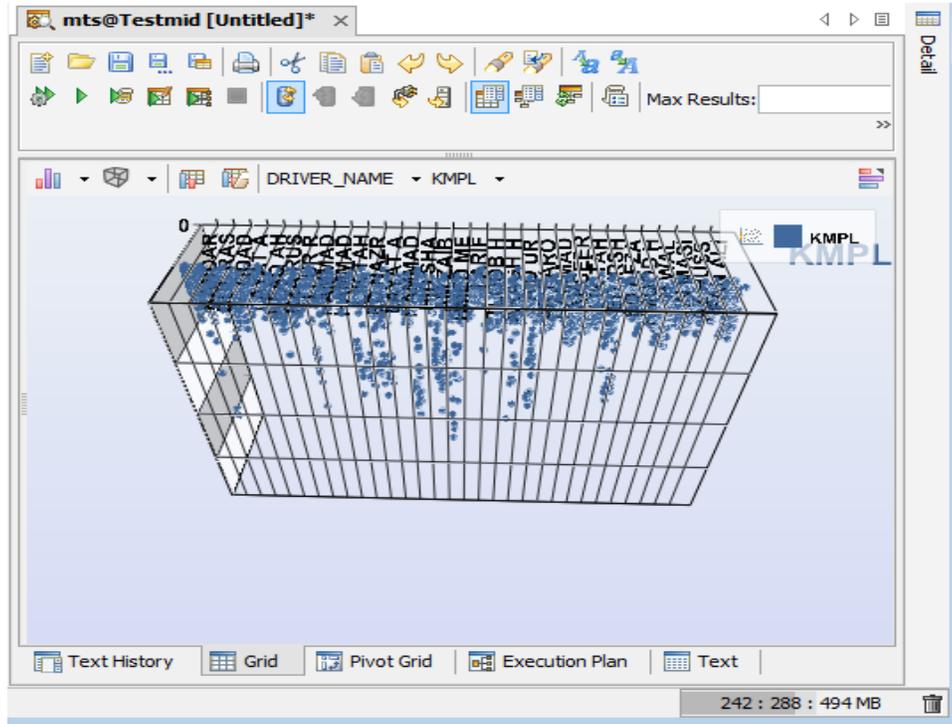


Fig. 12: Cluster analysis on the basis of drivers

In Fig. 12, it is very much clear that large numbers of anomalies are present in drivers based data. Many drivers have an abnormal behavior in cluster analysis. These drivers include BSH, D UR, ZAIB, AH H, D ME, SHA and AZAR are the main drivers whose data has an abnormal behavior and marked as anomalies.

CONCLUSION

In traditional ways we need to examine all the values in data set for anomaly detection. Before analyzing we cannot mark any value as correct and cannot identify any value as anomaly.

From evaluation results of proposed system, It is concluded that before clustering process, anomaly %age calculation is giving better result as compare to clustering of original values. After %age calculation at least 50% values can be marked correct and then we need to investigate anomalies in remaining 50% values. In this way we need 50% less effort to identify anomalies in data set.

The analysis of the data in all respect shows that our anomaly calculation and detection algorithm give a better result in terms of time constraints, resource utilization and variety of anomalies detection.

REFERENCES

Annie, G., 2012. Anomaly detection based on machine learning: Dimensionality reduction using PCA and classification using SVM. *Int. J. Comp. Appl.*, 47(21): 5-8.

Ashoor, A.S. and S. Gore, 2011. Importance of Intrusion Detection System (IDS). *Int. J. Scient. Eng. Res.*, 2(1): 1-4.

Basha, M.A., 2012. A simplified approach to agent based efficient anomaly intrusion detection in ad-hoc networks using honey tokens. *Int. J. Eng. Res. Appl.*, 2(3): 2242-2247.

Janakiraman, S. and V. Vasudevan, 2009. An intelligent distributed intrusion detection system using genetic algorithm. *J. Converg. Inform. Technol.*, 4(1): 70-76.

Pari, S.N. and D. Sridharan, 2011. A performance comparison and evaluation of analysing node misbehaviour in MANET using intrusion detection system. *Int. J. Comp. Sci. Emer. Technol.*, 1(1): 35-40.

Ramos-Morillo, H., I. Lorenzo-Fonseca, F. Mora-Gimeno, D. Marcos-Jorquera and J.A. Gil-Martínez-Abarca, 2010. Network intrusion detection system embedded on a smart sensor. *IEEE Trans. Indus. Elec.*, 58(3): 722-732.

Ringberg, H., A. Soule, J. Rexford and C. Diot, 2007. Sensitivity of PCA for traffic anomaly detection. *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, San Diego, California, USA.

Shah, B. and B.H. Trivedi, 2012. Artificial neural network based intrusion detection system: A survey. *Int. J. Comp. Appl.*, 39(6): 13-18.

- Snelding, K., 2010. Audit of East Service Center Employee Fuel Theft, Report.
- Thiprungsri, S., 2010. Cluster analysis for anomaly detection in accounting data. Proceedings of the 19th Annual Strategic and Emerging Technologies Research Workshop, San Francisco, California, USA.
- Thottan, M. and J. Chuanyi, 2003. Anomaly detection in IP networks. *IEEE Trans. Signal Proc.*, 51(8): 2191-2204. doi: 10.1109/TSP.2003.814797.