Anomaly Detection in Fuel Expense And Mileage Through Data Mining

Kaleem Habib, Arif Iqbal Umar, Noor Ul Ameen, Muhammad Ali And Shoukat Mehmood

***Abstract*—** In large organizations a notable portion of the fuel budget is misused by malpractices of employees and fuel providers. An optimal usage of this amount could be of big advantages to the organization. We proposed a novel data clustering algorithm based on mode of the current data to determine the misuse of the fuel of the vehicles. The results reflect that this algorithm could be used to implement an effective check on the misuse of the fuel in big organizations.

***Keywords*—**data mining, fraud, fuel expense and mode.

I. Introduction

FUEL misuse (theft/fraud) is a hot problem. 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.

Kaleem Habib is a graduate student at Department of Information Technology, Hazara University Mansehra, Pakistan. (phone: +92 992 414140; fax: +92 992 414111; e-mail: kaleemhabib_2000@yahoo.com).

Arif Iqbal Umar is with Department of Information Technology, Hazara University Mansehra, Pakistan. (e-mail: arifiqbalumar@yahoo.com).

Noor ul Ameen is with Department of Information Technology, Hazara University Mansehra, Pakistan. (e-mail: nameen@hu.edu.pk).

Muhammad Ali is a graduate student at Dept. of Information Technology, Hazara University Mansehra, Pakistan. (e-mail: alibaba_ciit@yahoo.com).

Shoukat Mehmood is a graduate student at Dept. of Information Technology, Hazara University Mansehra, Pakistan. (e-mail: shoukat@ciit.net.pk)

We applied a novel cluster analysis algorithm to a dataset provided by a major logistic company in Pakistan to examine the use of mode of data set as a central point to group data into different clusters. Data set is comprised of the data reflecting Fuel Consumption Values (KMPL- Kilometer per Litter). These values are grouped into clusters of KMPL. The center of the cluster is mode of the fuel consumption values, which is present in the data for the most of times. The results reflect that the algorithm could be used to implement an effective check on the misuse of the fuel in the big organizations.

The rest of the paper is organized as follows. Section II presents the literature review. Section III describes the proposed algorithm. Section IV evaluates the results of the algorithm and Section V concludes the paper.

II. Literature Review

In the accounting literature, most studies focus on management fraud [1]. For the prediction of management fraud, most prediction models employ either logistic regression techniques or the Neural Network.

Anomaly detection in the field of Network Traffic, Network Security, Information Security, Node Behavior and wireless Network is mainly depend on some pre-defined characteristic, Historical data and different hardware settings [2]–[9]. Dimensionality Reduction and Classification methods 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 [6]. Historical data is used to verify the normal behavior of the system. The challenges are; User genuinely changes a system file, 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 [11]. The report observed the following:

1) The Fleet Services should install appropriate security devises to prevent or minimize theft of fuel form City owned vehicles. This could include the installation of locking fuel caps and/or anti-siphoning devises in City vehicles.
2) 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

& 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 [10]. The system is able to operate independently as an anomaly-based NIDS or integrated, transparently, in a Distributed Intrusion Detection System (DIDS).

A system is proposed for cluster analysis in [1], 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.

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. 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 vouches. 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 Detection Algorithm by introducing a mode as center point. The propose algorithm will work on a database level and highlight any suspicious entries for further investigation.

III. Proposed Solution

In this section a novel technique is proposed. The New technique will obtain a set of characteristic from current set of data and don't rely on pre define characteristics and Historical data because in case of financial transaction historical data become invalid.

For outliers, distance can be measure on the basis of neighborhood, on the basis of predefine center point. But this new technique will calculate a center point for each category of vehicle type on the basis of mode of each type of vehicle data set.

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

	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 possible 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	*Plot a data on the basis of distance **Di** from the center point **Ci** of **KPML***
	*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 points in step **2***
	1st Loop End
	}
	Algorithm Ends

IV. Evaluation

This study examined the possibility of using mode as center point to form clusters for fraud detection in fuel consumption. The algorithm is tested on a data set from a major logistic company in Pakistan. KMPL's value with maximum repetition which is a mode is used as center point. These tests will be performed on one selected vehicle categories. This study is a preliminary step to apply the cluster for fraud detection in automobile industry.

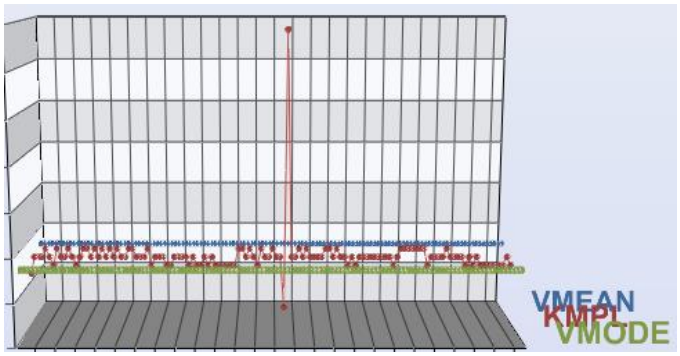

Fig. 2 Visualization of the Mode and Mean with KMPL

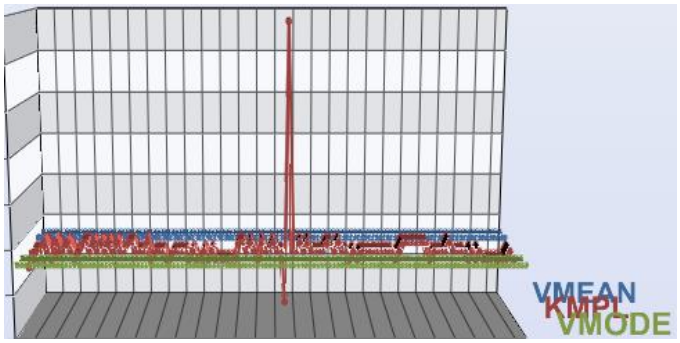

Fig. 3 Visualization of the Mode and Mean with KMPL

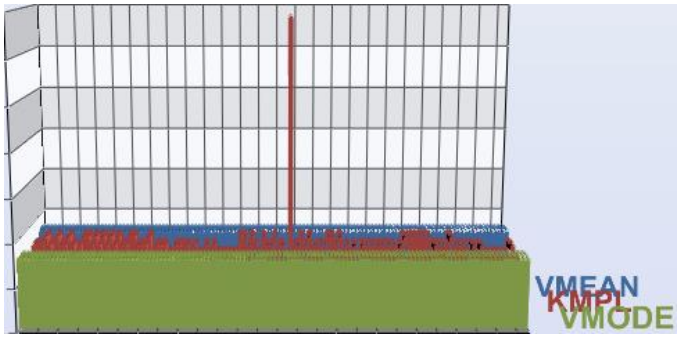

Fig. 4 Visualization of the Mode and Mean with KMPL

The analysis of the data in above three pictures in three different ways reveals that, it is very much clear that mean of data set is displayed above the maximum points of the original data set and if take mean as center point then it will give wrong result. In case of mode of data set, mode is overlapping maximum points of data set. It is concluded that mode is giving better result as compare to the mean and if we take mode as center point then it will give better result in clustering because it is a more similar to other points in data set.

V. Conclusion

In case of mean of data, a new data point is generated and all of the data set points are compared with new generated data point. If new generated data point in the form of mean become wrong then all data point compare with this point will also generate wrong result. In this case there is need for a data point which must exist in the existing data set. Therefore the mode of all data set is a good candidate for center point. From evaluation results it is clear that cluster made on the basis of mode as center point will be denser then cluster made on the basis of mean of data.

References

[1] S. Thiprungsri, "Cluster analysis for anomaly detection in accounting data," in *Proceeding of 19th Annual Strategic and Emerging Technologies Research Workshop* San Francisco, California, USA, 2010.

[2] B. Shah and B. H. Trivedi, "Artificial neural network based intrusion detection system: A survey," *International Journal of Computer Applications*, vol. 39, no. 6, 2012.

[3] M. Thottan and C. Ji, "Anomaly Detection in IP Networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[4] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," *ACM SIGMETRICS Performance Evaluation Review*, vol. 35, no. 1, pp. 109–120, 2007.

[5] S. Janakiraman and V. Vasudevan. "An intelligent distributed intrusion detection system using genetic algorithm," *Journal of Convergence Information Technology* vol. 4, no. 1, pp. 70–76, 2009.

[6] A. S. Ashoor and S. Gore. "Importance of Intrusion Detection system (IDS)," *International Journal of Scientific and Engineering Research* 2, no. 1, pp. 1–4, 2011.

[7] S. N. Pari and D. Sridharan. "A performance comparison and evaluation of analysing node misbehaviour in MANET using intrusion detection system," *International Journal of Computer Science Engineering and Technology*, vol. 1, no. 1, pp. 35–40, 2011.

[8] A. George, "Anomaly detection based on machine learning: Dimensionality reduction using PCA and classification using SVM," *International Journal of Computer Applications*, vol. 47, no. 21, 2012.

[9] M. A. Basha, "A simplified approach to agent based efficient anomaly intrusion detection in ad-hoc networks using honey tokens," *International Journal of Engineering Research and Applications*, vol. 2, no. 3, pp. 2242–2247, 2012.

[10] F. Maciá-Pérez, F. Mora-Gimeno, D. Marcos-Jorquera, J. A. Gil-Martínez-Abarca, H. Ramos-Morillo and I. Lorenzo-Fonseca, "Network intrusion detection system embedded on a smart sensor," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 3, 2010

[11] R. K. Snelding, "Audit of east service center employee fuel theft", Technical Report, no. CAO 2901-1011-04, City of Las Vegas, Nevada, 2010.