Credit Card Number Fraud Detection Using K-Means with Hidden Markov Method

Pooja Bhati[1#], Manoj Sharma[2*]
[1#](M.Tech Computer Science and Engineering, Rawal Institute of Engineering and Technology, India)
[2*](Assistant Professor, Rawal Institute of Engineering and Technology, India)

Abstract

Clustering is a way of segmenting the data into some purposeful groups. When done efficiently, the final product, i.e. the clusters should seize the very essence of the original data. Clustering and outlier detection are the two paramount fields of data mining. In today's time, when security is one of the major issues in every aspect of life, outlier detection becomes inevitable for data mining. Any arrangement or design which is contradictory to the rest of the arrangement can be defined as an outlier for that particular sample-set. Monetary fraud is hugely spread over every possible aspect of life. Credit card fraud is also very common, as they are being extensively nowadays. The method being proposed in this paper is to use k-means clustering and then finding outliers in the resultant clusters using the Hidden Markov Method. Our proposed algorithm effectively sub-divides the in-liners into clusters and then detects the outliers. After that Luhn Algorithm is being used for validating the resultant credit card numbers. Our proposed work would work much more efficiently and effectively in case of Big data, as normal k-means has a poor tendency to work with big data and also to detect outliers.

Keywords: - *Outlier, Monetary fraud, k-Means, Hidden Markov Method, Luhn Algorithm*

I. INTRODUCTION

The practice of thoroughly scrutinizing large data stores to assess the possibilities of finding potential structures or trends is called Data Mining. Data mining consists of some innovatory algorithms for separating data in a manner that is nothing like a simple analysis. Outlier detection is one of the influential and key issues in Data Mining. The definition for outlier given by Hawkins "An outlier is an observation which deviates so much from other observation as to arouse suspicions that it was generated by a different mechanism" can be quoted true for most of the cases.

Basically there are two major reasons for searching outliers: 1) to detect anomalies to make sure that our present data is not polluted 2) outliers are not necessarily representing corrupted or wrong data every time, there is a possibility that they have some hidden, unknown yet useful patterns in them. A typical example of applications focusing on detecting outlier is security related applications where the system, assuring the security, check for a predefined and similar pattern and on finding a tiny bit deviation from these patterns, it becomes alert. This statement is extremely true for the task we are doing, since our work of field is credit card numbers, for security is a major issue.

It is evident that performance of data mining algorithms and techniques can be amplify remarkably by a precise and well organized elimination of outliers. This approach of diagnosing and removing outliers is known as Data-Cleaning, when used in other applications as a pre-processing activity. Data mining has a no. of different regions or territories which are extensively using outlier detection, which in turn results in a massive and highly distant literature for methods of outlier detection. Some of these methods can be used for more general problems while others can be used for specific problems. Many applications such as network sensors, intrusion detection, fraud detection and those for marketing etc. ar using outlier detection as their inevitable part. Many researchers argued about whether using clustering techniques for outlier detection is feasible or not.

In our proposed work, we are using k-means clustering technique with Hidden Markov Method (HMM) along with Luhn algorithm for outlier detection and validation in big data consisting credit card numbers. Basically this methodology is combining K-Means and HMM for clustering and outlier detection with Luhn algorithm for validity checks.

II. METHODOLOGIES USED

A. K-Means Clustering

Clustering techniques are being used for fulfilling the purpose of finding new and substantial information from a given data-set by going through it. We can find a large number of clustering techniques for a hugely diversified real life work fields. Among all those clustering methods K-means can be easily denoted as of the earliest methods to be formed. It can also be viewed as one of the simplest yet effective techniques for clustering. One can easily understand the working of K-means better by looking at its general algorithm. The algorithm has following step:-

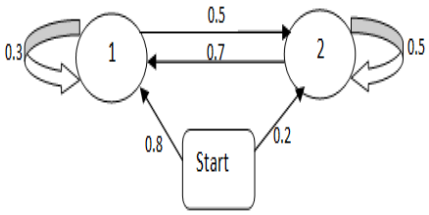
1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.

B. Hidden Markov Model

The basic theory of Hidden Markov Model (HMM) was first published in a classic papers series by Baum and his co-workers in the end of 1960's and beginning of 1970's. But the implementation of HMM for speech processing application was done by Baker in the 1970's. However it's only the recent past years when a thorough understanding and application of its theory has occurred.

A Hidden Markov Model can be described as a statistical Markov Model where a system is being modeled which is assumed to be a process (Markov) having hidden or unobserved states. A simplest dynamic Bayesian Network can be representation of HMM. L.E.Baum and his co-workers developed the mathematics behind the HMM.

Here's an example for better understanding of HMM.



Initial probabilities $\pi = (0.8, 0.2)$ T

Transition probabilities

$M = (mij: = P(i \rightarrow j))\ i, j=1, 2$

$$\begin{pmatrix} 0.3,0.7 \\ 0.5,0.5 \end{pmatrix}$$

Emission probabilities,

e.g. e1b= 0.5, e2c= 0.45.

Random source (Xt) with values in

$\Sigma = \{a, b, c\}$:

e.g.: $PX(X1= a, X2= b) = \pi 1e1a\ (a11e1b+ a12e2b) + \pi 2e2a\ (a21e1b+ a22e2b)$

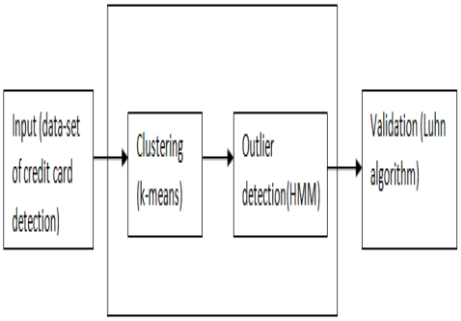
***Luhn Algorithm*: -**Luhn algorithm has been named after an IBM scientist Hans Peter Luhn. The sheer use of this algorithm is to generate and validate Credit card no. Another widely used name for this algorithm is "modulus 10" or "Mod 10" algorithm. Mod 10 is a simple checksum formula and can easily be used to validate a range of identification numbers, which can be credit card numbers, IMET number etc. This algorithm is being used exhaustively today and can be find in public domain. The original design of this algorithm was made to protect against accidental errors and not malicious attacks.

The formula works by appending a check digit to a partial account no. to generate a complete account no. further this no. must pass the following test:-

1. Starting from the rightmost digit (also the check digit), double every second digit; if the sum is more than 9 then add the digits (e.g. 16, 1+6=7).
2. Take the sum of the complete number.
3. If the total modulo 10 is equal to 0 then the no. is valid; else it is not valid.

III. PROPOSED SYSTEM

Figure below is giving architecture of the desired system. Firstly an input, ideally some dataset, will be given. Then data is divided into different clusters. After that fraud detection is done. And finally validation check is applied.



Input:-An excel file consisting of credit card numbers stored in a database imported to our project and provides use the input values.

Clustering process:-K-Means is one of most popular and easy widely used clustering technique for the purpose of grouping data objects in groups or clusters. Clustering is very useful for reducing the size of bulky datasets. Firstly, centroids for each group are calculated or chosen randomly. And then data with similar characteristics i.e. closer to its respective centroids are grouped together.

Outlier Detection: - Hidden Markov Model is a statistical model which essentially works with Markov model, where we have unobserved or hidden states. Hidden Markov Model is being used exhaustively in the field of artificial intelligence and more specifically in speech recognition. In our proposed system HMM is doing the job of detecting

outlying objects from the final resultant data. Thus reducing the size of the dataset and eliminating the erroneous data.

Luhn Algorithm: - Once the dataset of credit card numbers has been clustered and outlier detection has been successfully done by using K-Means and HMM respectively, the process of validation of the credit card no. would be started using the Luhn algorithm. On applying Luhn algorithm, the list of those credit card numbers which had been generated by some generation formula is obtained and those are the valid no.

Algorithm: -

1. K-Means, using nearest neighbor method for calculating centroids, is applied on the data-set.
2. Credit card no. are clustered into groups.
3. After that Hidden Markov Model is used for detecting the outlying credit card no. from the clusters.
4. Clusters having outlying credit card no. are hidden are rest of the clusters are shown as correct credit no.
5. In the next step, Luhn algorithm is applied to checking the credit card no. whether they had been generated by the Luhn algorithm or not.
6. On fulfilling the condition the no. are valid, otherwise invalid.

IV. EXPERIMENTAL RESULTS

The system that has been generated by the proposed work creates an interface, where we can easily perform several tasks by clicking on tabs. By providing an interface this system has become quite user-friendly and easily understandable. Following screen-shots can help to understand how system is working and working of different-different tabs.

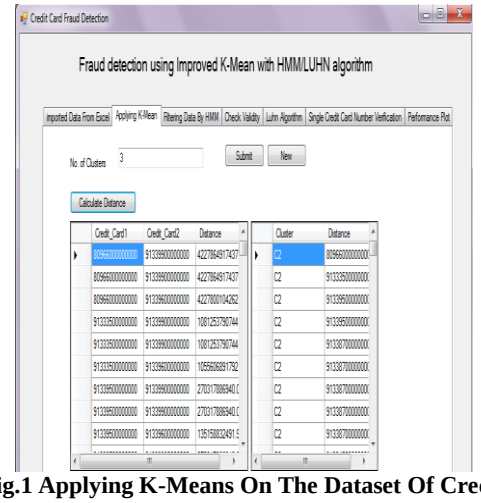


Fig.1 Applying K-Means On The Dataset Of Credit Card No.

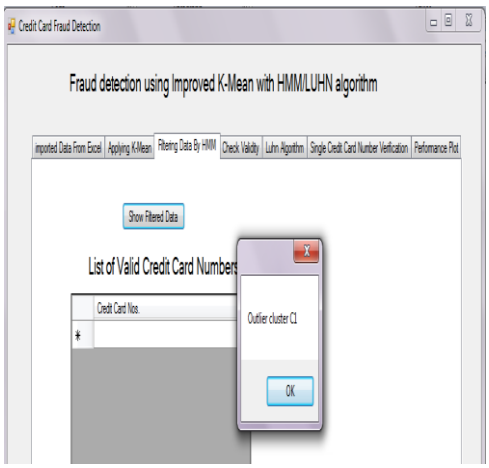


Fig.2 Outlier Detection Using HMM

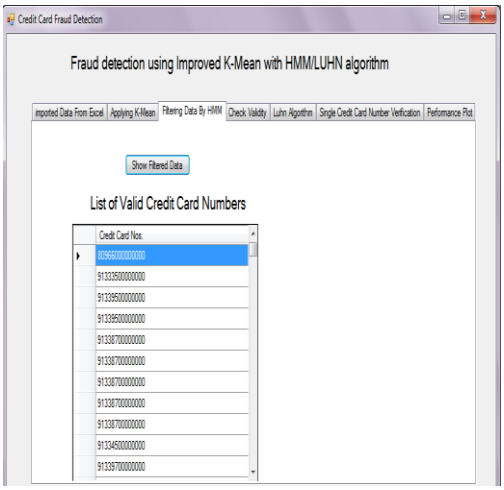


Fig.3 Filtered Data List After Applying HMM

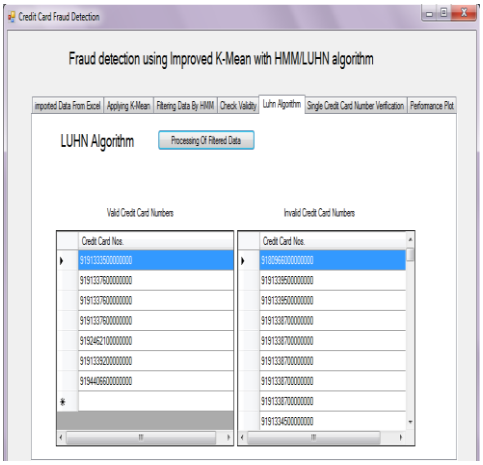


Fig.4 Checking Validation Of Credit Card No. Using Luhn Algorithm

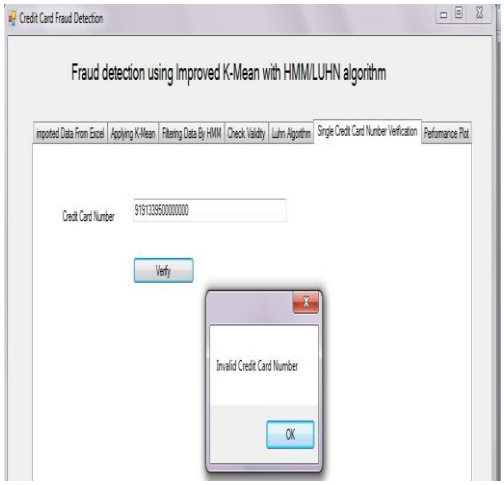


Fig.5 Checking For Valid Or Invalid No.

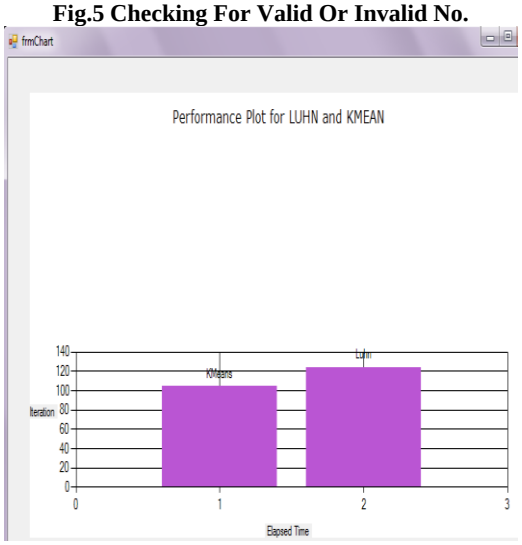


Fig.6 Comparative performance plot for Luhn and K-Means.

V. CONCLUSION

On completion of the underlying system we can conclude that the system is providing far better system performance efficiency than a system using k-means for outlier detection. Since our main focus is on finding fraudulent data in a database of credit cards hence efficiency is measured on the basis of frequency of detecting outliers or wrong credit card no. For this purpose our system has a mechanism consisting of K-Nearest neighbor algorithm with Hidden Markov Model. After getting a refined list of credit cards Luhn is used for checking if the resultant credit card no. is valid or not. Because of using KNN and HMM before LUHN algorithm, it gives better results than the lone application of Luhn algorithm. So we are having a system which is efficiently detecting wrong/fraudulent credit card no. as a final product.

The future scope for this system can be working with more attributes of the credit card no. As the technology is growing rapidly hackers are finding new ways to crack the security means, so by working with more attributes we can make the system more complex. This in turns will make the system safer.

REFRENCES

[1] Credit Card Fraud Detection by Improving K-Means, Mahesh Singh, Aashima, Sangeeta Raheja, International Journal of Engineering and Technical Research (IJETR), ISSN: 2321-0869, Volume-2, Issue-5, May 2014

[2] Clustering Memes in social media streams, Mohsen JafariAsbagh, Emilio Ferrara, Onur Varol, (published online: 18 November, 2014.)

[3] A hybrid network intrusion detection framework based on random forests and weighted k-means. Reda M. Elbasiony, , Elsayed A. Sallam1, , Tarek E. Eltobely2, , Mahmoud M. Fahmy3 (Ain Shams Engineering Journal Volume 4, Issue 4, December 2013).

[4] Enhance Luhn Algorithm for Validation of Credit Cards Numbers , Khalid Waleed Hussein , Dr. Nor Fazlida Mohd. Sani , Professor Dr. Ramlan Mahmod , Dr. Mohd. Taufik Abdullah , IJCSMC, Vol. 2, Issue. 7, July 2013, pg.262 – 272

[5] A comparative study of efficient initialization methods for the k-means clustering algorithm M. Emre Celebi, Hassan A. Kingravi , Patricio A. Vela , Expert Systems with Applications, Volume 40, Issue 1, January 2013, Pages 200–210.

[6] Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach, Ms. S. D. Pachgade, Ms. S. S. Dhande, Volume 2, Issue 6, June 2012, International Journal of Advanced Research in Computer Science and Software Engineering.

[7] A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining, H.S.Behera Abhishek Ghosh, Sipak ku. Mishra, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.

[8] Regularized k-means clustering of high-dimensional data and its asymptotic consistency
Wei Sun, Junhui Wang, and Yixin Fang, Electron. J. Statist. Volume 6 (2012), 148-167.

[9] An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification, Sanjay Kumar Pankaj Pandey ; Susheel Kumar Tiwari ; Mahendra Singh Sisodia, Advances in Engineering, Science and Management (ICAESM), 2012 International Conference

[10] The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature E.W.T. Ngai , Yong Hu , Y.H. Wong , Yijun Chen, Xin Sun, Decision Support Systems Volume 50, Issue 3, February 2011, Pages 559–569

[11] A Hidden Markov Model Based Method for Anomaly* Detection of Precipitation Series by Jun Shen, Minhua Yang, Ronghua Zhong, Cuchai Zhang, Journal of Information & Computational Science 8: 9 (2011) 1551–1560

[12] Equations for Hidden Markov Model, Alexander Schonhuth, 2008.

[13] Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

[14] Outlier Detection in Clustering, Svetlana Cherednichenko, 24.01.2005, University of Joensuu, Department of Computer Science, Master's Thesis.