ISSN (Online) : 0974-5645
ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(2), 165–171, January 2015* DOI : 10.17485/ijst/2015/v8i2/58081

An Optimized Multiple Semi-Hidden Markov Model for Credit Card Fraud Detection

A. Prakash[1] and C. Chandrasekar[2]

Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India; rakashphd789@gmail.com
Department of Computer Science, Periyar University, Salem, Tamil Nadu, India

Abstract

The credit card payment system is a widespread usable system which provides the easiest way of payment to the customers, but some of them misuse another individual's credit card for personal reasons. So, in order to provide credit card fraud detection, Multiple Semi-Hidden Markov Model is suggested to gather multiple observations and the detection phase is executed. It is significant to compute the good model parameters because it impacts the detection performance in the Multiple Semi-Hidden Markov Model. So, in this manuscript an innovative technique is introduced which is called Optimized Multiple Semi-Hidden Markov Model (OMSHMM) which is used for optimizing the model parameters. The Multiple Semi-Hidden Markov Model is used for detecting fraudulent users and for optimizing training values Cuckoo Search algorithm is proposed. The main intent of this research is automating the use of Multiple Semi-Hidden Markov Model, by liberating customers from the necessity of statistical knowledge. The number of states and also its model parameters are decided by the Cuckoo Search algorithm. An experimental result shows that when compared to the existing research there is high accuracy in the proposed research.

Keywords: Credit Card Fraud Detection, Cuckoo Search Optimization, Multiple Semi-Hidden Markov Model

1. Introduction

Recently, credit card usage has increased among the customers. Credit card is used in online or it may be used in ordinary shopping. Because, credit card payment is an essential one and it is expedient to utilize[1]. Because of the increase and rapid enlargement in E-Commerce, the utilization of credit cards for purchasing has also enlarged radically. As the usage of credit cards are increases, the cases of credit card fraud related with it is also increasing. The credit card fraud is defined as an illegal activity by a customer for whom the account was not anticipated[2]. The customers who are utilizing the credit card has not all having the connection with the cardholder and has no intention of making the repayments for the procure they done. So, to examine the spending behavior of the customers the fraudulent persons are detected. Detecting the fraud is defined as recognizing the suspicious users[3,4]. If there is any irregularity happens in the spending actions then it is considered as suspicious and taken for further deliberation.

There are numerous techniques[5] are suggested for the credit card fraud detection. For the accomplishment of the fraud management systems several techniques such as rule-based scoring method[6], statistical tools and artificial neural networks are presented. Scoring model is defined as a set of policies which illustrates the distinctiveness of some particular known fraud categories. If the fraud categories are simple and manageable scoring model is effective. Least-squares regression method is one of the statistical investigation methods[7] which are used to detect the fraud audits. In the credit card fraud recognition, the Hidden Markov Model is utilized which has circled out to be one of the superior part successful techniques in the area of assessment and recognition[8]. In the traditional Hidden Markov Model method, move

*Author for correspondence

toward the state duration is moreover of a unit interval or completely undetermined to be geometrically isolated to create the primary method Markovian. Semi-Hidden Markov Model is an extension of Hiden Markov Model which is designed to permit all-purpose allotment for the state intervals.

The major drawback is takes place in reciprocally Hidden Markov Model and Semi-Hidden Markov Model i.e., it is normally anticipated that there endures at least one observation associated with each and every state in which the Hidden Markov chain acquires on. To develop the usefulness of the Semi-Hidden Markov Model to assimilate the multiple observation of Semi-Hidden Markov Model which is named as Multiple Semi-Hidden Markov Model (MSHMM)[8]. By utilizing this detection accuracy is improved which is superior to Semi-Hidden Markov Model. The major drawback in this method is to compute the model parameters is significant which impacts the performance of the detection accuracy. So, a novel technique is introduced named as Optimized Multiple Semi-Hidden Markov Model (OMSHMM) for optimizing the model parameters. For optimization, Cuckoo Search algorithm is proposed. By using this algorithm, the number of states, and probability distributions of states are decided.

The remainder of this work is as follows: In section 2 the previous research on credit card detection methods are discussed. In Section 3 the Multiple Semi-Hidden Markov Model is described. In Section 4 Proposed work is presented. Section 5 deals the performance analysis. Section 6 presents the conclusion and future work.

2. Previous Research on Credit Card Fraud Detection Methods

In this section, some of the previous research about the credit card fraud detection methods is suggested.

In[9] Shawe-Taylor et al. suggested a scoring method which is set of policies that illustrates the features of the specific fraud categories. This scoring method is efficient if the fraud categories are trouble-free and convenient in the scale and complication of rule parameters. Even if the fraud type is simple it is necessary to define the entire policies. Furthermore, this scoring method is not self-learning ability and this method is not able to rapidly acclimatize to new fraud categories.

In[10] Huang et al. suggested a hybrid financial investigation method which includes static and trend examination methods to create and train a back-propagation neural network model. This method is capable to give high prediction accuracy and also performs other methods like discriminate analyses, decision trees, and back-propagation neural networks alone. On the other hand, pure statistical tools can scarcely make available an adaptive learning capacity in order to handle the fraud recognition.

Ekrem Duman et al. presented to develop a technique[11] which enhances the detection rate of fraudulent in credit card transaction. In this technique, the entire transaction is scored. According to the scores the whole transaction are sort out as fraudulent or lawful users. The main intent is to diminish the incorrectly sort out number of transactions in the credit card fraud detection system. Additionally, wrong category of the transaction does not have the identical collision in that when a card in the hand of fraudsters its complete available limit is utilized. Consequently the misclassification cost should be considered as the available limit of the card.

Quah and Srinagesh et al. suggest a framework[12] which can be used in real time applications in which firstly an outlier examination is made independently for each and every customer using self organizing maps after that an analytical method is utilized to categorize the irregular looking transactions. In[13] Panigrahi et al. suggests a four component fraud recognition solution which is associated in a serial manner. The main intent is to decide a set of suspicious transactions and by utilizing the Bayesian learning method on this list to envisage the frauds. In[14] Sanchez et al. presented a different method and it is used association rule mining to classify the patterns for normal card usage and denoting the ones not fitting to these patterns as suspicious.

In[15] Wen-Fang Yu et al. suggested an outlier mining technique to discover the credit card frauds. For outlier mining method, the distance based techniques are used. This method efficiently recognizes the overdrafts and also to visualize the deceptive transactions. You Fucheng et al.[16] suggested the behavior model for the credit card fraud recognition system. For identifying the fraud, the chronological behavior analysis is utilized. By make use of the transaction record of a single credit card, the model is produced. In this method, unsupervised self organizing map method is utilized to identify from the normal ones.

In[17] Chuang et al. presented a model according to the data mining. This method uses the web services to replace data between banks and fraud pattern mining algorithm

for recognition. In this scheme, the participant banks can distribute the information about fraud patterns in a heterogeneous and disseminated environment and further improve their fraud detection capacity and diminish economic loss.

Bayesian networks are also one of the methods which are used to recognize fraud in the credit card system[18]. By using this method, there are better results to recognize the fraud users. The main intent of a similarity tree by utilizing decision tree logic has suggested in[19]. The decision tree is defined recursively; it comprises nodes and edges which are sticked with attribute names and with values of attributes correspondingly.

In[20] Abhinav Srivastava et al. developed the Hidden Markov Model to recognize the fraudulent in the credit card recognition system. The Hidden Markov Model is originally trained with the normal activities of the cardholder. If the current transaction is not recognized by the trained HMM with high probability, it is determined as to be fake users. Anna Strizhak et al.[21] employed self organizing map method to create a fraud detection model. M. Huang et al.[22] suggested the online questionnaire technique to gather with the train data and the questionnaire responded transaction is utilized to predict the new transactions.

3. Multiple Semi-Hidden Markov Model Credit Card Fraud Detection

In the credit card detection system, multiple Semi-Hidden Markov Mode[l8] is suggested for identifying the illegal customers. For this model, consider a Markov chain with N states which are labeled as $\{1, 2, \ldots M\}$, in which the probability of transition from state m' to state m is represented $t_{m, m'}$, where $m, m' = 1, 2, \ldots N$, and the initial state probability distribution is represented by $\{\pi_m\}$. Let us consider a_t represent the state in which the system takes at time$t$, where $t = 1, 2, \ldots T$. The state sequence is denoted as $\{a_t\}$, but when desire to be clear about the interval, we approve the notation a_b^c, meaning $\{a_t: b \le t \le c\}$.

Correspondingly, g_t represents the observable output at time t related with state a_t, and let us consider the $b_m(a_t)$ be the probability of observing o_t, it is specified as $a_t = m$. Assume the conditional independence of outputs so that $b_m(g) = \prod_{t=b}^{c} b_m(g_t)$, in which g_b^c denotes the observation sequence from time b to c.

Assume the two sequences $\{g_t\}$ and $\{q_t\}$ are accessible as the outcomes of a Hidden Markov Model state sequence. In this model, the conditional probability which g_t emerges if the state is at m is denoted as $b_m(g_t)$ and the equivalent conditional probability for the second output is denoted as $c_m(q_t)$. Some random delay T is introduced between the two output sequences; these two sequences are no longer synchronized. The missed observation is denoted as the symbol ϕ_t which is the output at time t. Since the observation may not essentially be made at each and every time interval. The set of the observation time instants are represented by $G = \{t_1, t_2, \ldots t_n\}$, in which $1 \le t_1, t_n \le T$. The multiple sequences of observations are accessible. These multiple observation sequences may have their observation intervals, initial points which is different from others. The two observation sequences are denoted as $\{g_t\}$ and $\{q_t\}$. There is a delay T between the two observation sequences, where T takes on a value from $\{0, \pm 1, \pm 2..\}$. Either of the two streams can be any category of the observation or missing patterns analyzed.

4. Optimized Multiple Semi-Hidden Markov Model Credit Card Fraud Detection

The optimized Multiple Semi-Hidden Markov Models is proposed in order to optimize the model parameters. Because it is significant to optimize the model parameters so that to improve the accuracy in the detection performance. The MSHMM model parameters are decided during an iterative process which is called training phase. Cuckoo search optimization techniques can be utilized to optimize the Multiple Semi-Hidden Markov Model parameters.

Cuckoo search is an optimization algorithm which was motivated by the obligate brood parasitism. In this algorithm some of the cuckoo types used the nests of other birds to lay their eggs. This can be utilized for different optimization troubles. In this algorithm, each egg in a nest denotes the solution and new solution is a cuckoo egg. The most significant concept of this algorithm is to make utilize the new and potentially best solutions to replace a less good solution in the nests. This method can be enlarged to problematical situations in which every nest has multiple eggs which represent the set of solutions.

The major contribution of this research is:

- Gathering of the card holder's information and sustain in a central data base.
- Using cuckoo search optimization algorithm for optimizing the number of states and the model parameters of the Multiple Semi-Hidden Markov Model.
- Create the observation model with multiple observations.
- Clustering process is done for considering only the amount of the card holder.
- In the testing phase, the fraud customers are identified by comparing the probability values which is obtained in training and testing phase.

4.1 A Novel Algorithm for Optimized Multiple Semi-Hidden Markov Model

1. Collect card holder information.
2. // Training phase.
3. // MSHMM parameters.
4. Number of states (N) in the MSHMM.
5. Number of observations (M) in the MSHMM.
6. Transition probability (t) in the MSHMM.
7. Initial probability of the state (π_i) to start at time t = 1.
8. //Evolution of MSHMM parameters using Cuckoo search algorithm.
9. Initialize objective function f(x), x = (x_1, x_2, … x_d).
10. Create an initial population of n host nests
11. While (t<max Generation).
12. Get a cuckoo random and restore its solution by performing Levy flights.
13. Compute the fitness function F_i.
14. Fitness is calculated by computing the objective function,
15. $F_i \rightarrow P(O \mid \lambda) = \sum_{i=1}^{N} a_T(\mathrm{i})$, where $\lambda = N, M, T, \pi_i$
16. If ($F_i > F_j$).
17. Replace j by the new solution.
18. End if.
19. Identify the current best.
20. Pass the current best solutions to the next generation.
21. End while.
22. Take the optimum number of states and the probability value P_{tr} in the training phase.
23. Generate the observation model with multiple observations.
24. // Testing phase.
25. Obtain the amount of the cardholder via clustering process.
26. Obtain the probability value P_{te}.
27. If ($P_{tr} = P_{te}$).
28. No anomaly.
29. Else.
30. Detect the anomaly.
31. End if.

In the above algorithm the Cuckoo search optimization technique is used for optimizing the parameters in the Multiple Semi-Hidden Markov Model. Initially the information about the card holder is gathered. The observation symbols regarding to the particular card holder is identified dynamically. For this the clustering algorithm is executed on the past transactions. Usually, the user transactions are accumulated in the bank's database which includes many attributes. For the credit card detection system, only amount that the cardholder spent in his transaction is considered. The normal spending behavior of the customer is taken for the spending profile of a cardholder. There are three categories of the card holder according to their spending behavior. It is called high-spending group, medium-spending group, and low-spending group. At the end of the clustering method, the spending behavior of the cardholders is concluded.

In the training phase the Multiple Semi-Hidden Markov Model parameters like number of states, number of observations, transition probability, and initial probability of the state are optimized by using the cuckoo search optimization algorithm. In this algorithm, initialize the objective function and the population of the host nests. The fitness value is computed for every nest. Each and every solution is a complete Multiple Semi-Hidden Markov Model, so that the system computes how well the model is with respect to the observation sequence, so that evaluate its probability and allocate this value as its fitness.

In the resulting process, some of the nests will be replaced by new nests and also sustains the total population of constant size across generations and parents with the best fitnesses. This process is continued until the cycle obtains the maximum number of generations. So, by using this algorithm the optimum number of states and the probability value are obtained. After that to identify the illegal customers the probability value is compared which

is acquired from the training and the testing phase. If the two values are matched, there is no anomaly. Suppose, if there is mismatch between these values consider as an anomaly.

5. Performance Evaluation

In this section, the performance is validated to compare the existing Multiple Semi-Hidden Markov Model (MSHMM) and the proposed Optimized Multiple Semi-Hidden Markov Model (OMSHMM). The performance is validated in terms of precision, Recall, F-Measure. Based on the comparison and the results from the experiment shows the proposed work works better than the other existing systems with higher rate.

5.1 Precision

Precision value is evaluated according to the retrieval of information at true positive prediction, false positive.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

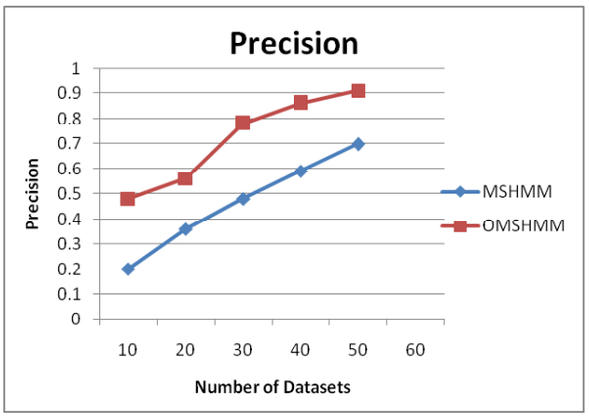

Figure 1. Precision.

Table 1. Precision vs. Number of Datasets

SNO	Number of Dataset	MSHMM	OMSHMM
1	10	0.2	0.48
2	20	0.36	0.56
3	30	0.48	0.78
4	40	0.59	0.86
5	50	0.698	0.91

The two parameters which are called Number of Dataset and precision are taken to compare the existing and the proposed system. In the X-axis Number of data sets are taken. In the Y-axis Precision is taken. This graph clearly shows that when compared to the existing method the precision is higher in the proposed method.

5.2 Recall

Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$Recall = \frac{True\ Positive}{(True\ positive + False\ negative)}$$

The two parameters which are called Number of Dataset and recall are taken to compare the existing and the proposed system. In the X-axis Number of data sets are taken. In the Y-axis Recall is taken. This graph clearly shows that when compared to the existing method the recall is higher in the proposed method.

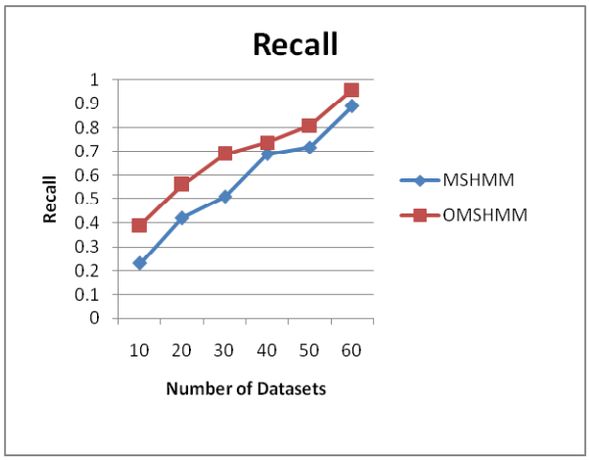

Figure 2. Recall.

Table 2. Recall vs. Number of Datasets

SNO	Number of Dataset	MSHMM	OMSHMM
1	10	0.23	0.39
2	20	0.42	0.56
3	30	0.51	0.69
4	40	0.69	0.74
5	50	0.72	0.81

5.3 F-Measure

F-Measure is a computation of a test's accuracy. The F-Measure score can be interpreted as a weighted average of the precision and recall.

F-Measure = 2.Precision.recall /(precision + recall).

The two parameters which are called Number of Dataset and F-Measure are taken to compare the existing and the proposed system. In the X-axis Number of data sets are taken. In the Y-axis F-Measure is taken. This graph clearly shows that when compared to the existing method the F-Measure is higher in the proposed method.

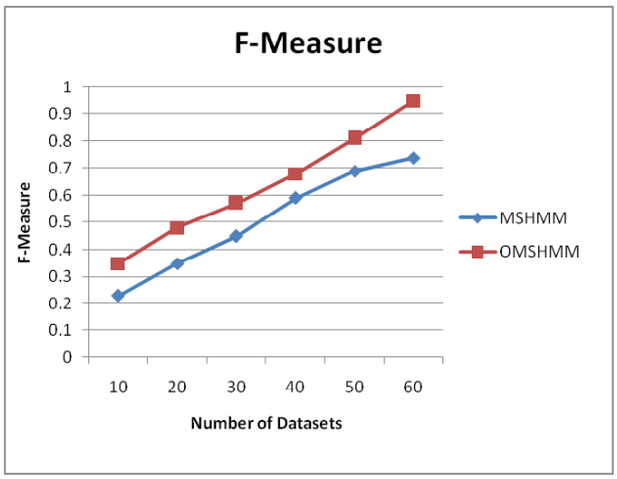

Figure 3. F-Measure.

Table 3. F-Measure vs. Number of genes

SNO	Number of Dataset	MSHMM	OMSHMM
1	10	0.23	0.35
2	20	0.35	0.48
3	30	0.45	0.57
4	40	0.59	0.68
5	50	0.69	0.81

6. Conclusion

Credit card fraud detection system is a significant one in order to reduce the mal-transaction. In the existing research, the Multiple Semi-Hidden Markov Model is used to collect the multiple observations and by using these observations the fraudulent customers are detected. But optimizing the model parameters of the Multiple Semi-Hidden Markov Model is an important one to improve the accuracy. So, in this work Optimized Multiple Semi-Hidden Markov Model is proposed to optimize the model parameters so that very effectively detect the anomaly. For future work, it is must to develop a system which control credit card fraud before any real transaction is made.

7. References

1. Dheepa V, Dhanapal R. Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft Computing. 2012; 4(4):391–7.
2. Patidar R, Sharma L. Credit card fraud detection using neural network. International Journal of Soft Computing and Engineering (IJSCE). 2011; 1(NCAI2011):32–8.
3. Allan T, Zhan J. Towards fraud detection methodologies. IEEE Proceedings of the Fifth International Conference on Future Technology (Future Tech); 2010.
4. Dhanapal R. An intelligent information retrieval agent. Knowl Base Syst. 2008; 21(6):466–70.
5. Quah TS, Sriganesh M. Real-time credit card fraud detection using computational intelligence. Expert Syst Appl. 2008; 35(4):1721–32 .
6. Viaene S, Ayuso M, Guillen M, Gheel DV, Dedene G. Strategies for detecting fraudulent claims in the automobile insurance industry. Eur J Oper Res. 2007; 176:565–83.
7. Mercer CJ. Fraud detection via regression analysis. Comput Secur. 1990; 9(4):331–8.
8. Prakash A, Chandrasekar C. An ensemble approach for credit card fraud detection. Int J Comput Appl. 2012; 59(19):1–6.
9. Shawe-Taylor J, Howker K, Burge P. Detection of fraud in mobile telecommunications. Inform Secur Tech Rep. 1999; 4(1):16–28.
10. Huang S, Tsai C, Yen C, Cheng Y. A hybrid financial analysis model for business failure prediction. Expert Syst Appl. 2008; 35(3):1034–40.
11. Duman E, Ozcelik MH. Detecting credit card fraud by genetic algorithm and scatter search. Journal of Expert Systems with Applications. 2011; 38(10):13057–63.
12. Quah JTS, Srinagesh M. Real-time credit fraud detection using computational intelligence. Expert Syst Appl. 2008; 35(4):1721–32.
13. Panigrahi S, Kundu A, Sural S, Majumdar A. Credit card fraud detection a fusion approach using Dempster-Shafer theory and Bayesian learning. Inform Fusion. 2009; 10(4):354–63.
14. Sanchez D, Vila MA, Cerda L, Serrano JM. Association rules applied to credit card fraud detection. Expert Syst Appl. 2009; 36(2):3630–40.

15. Yu W-F, Wang N. Research on credit card fraud detection model based on distance sum. Proceedings of the International Joint Conference on Artificial Intelligence; 2009. p. 353–6.
16. Zhang Y, You F, Liu H. Behavior-based credit card fraud detection model. Fifth International Joint Conference on INC, IMS and IDC (NCM '09); 2009. p. 855–8.
17. Aleskerov E, Freisleben B, Rao B. CARDWATCH: a neural network based database mining system for credit card fraud detection. Proceedings of the IEEE/IAFE on Computational Intelligence for Financial Engineering; 1997. p. 220–6.
18. Maes S, Karl T, Bram V, Bernard M. Credit card fraud detection using Bayesian and neural networks. Proceedings of 1st NAISO Congress on Neuro Fuzzy Technologies. Hawana; 2002.
19. Kokkinaki AI. On atypical database transactions: identification of probable frauds using machine learning for user profiling. IEEE Knowledge and Data Engineering Exchange Workshop (KDEX); 1997. p. 107.
20. Srivastava A, Kundu A, Sural S, Majumadar AK. Credit card fraud detection using Hidden Markov Model. IEEE Transactions on Dependable and Secure Computing. 2008; 5(1):37–48.
21. Zaslavsky V, Strizhak A. Credit card fraud detection using self organizing maps. Int J Inform Secur. 2006; 18:48–63.
22. Chen R, Chiu M, Huang Y, Chen L. Detecting credit card fraud by using questionnaire-responded transaction model based on Support Vector Machines. Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning; 2004. p. 800–6.