Use of Possibilistic Fuzzy C-means Clustering for Telecom Fraud Detection

Sharmila Subudhi and Suvasini Panigrahi

Abstract This paper presents a novel approach for detecting fraudulent activities in mobile telecommunication networks by using a possibilistic fuzzy c-means clustering. Initially, the optimal values of the clustering parameters are estimated experimentally. The behavioral profile modelling of subscribers is then done by applying the clustering algorithm on two relevant call features selected from the subscriber's historical call records. Any symptoms of intrusive activities are detected by comparing the most recent calling activity with their normal profile. A new calling instance is identified as malicious when its distance measured from the profile cluster centers exceeds a preset threshold. The effectiveness of our system is justified by carrying out large-scale experiments on a real-world dataset.

Keywords Call detail records · Clustering · Possibilistic fuzzy c-means · Fraud detection

1 Introduction

In recent years, the usage of mobile phones for communication has revolutionized the telecom industry along with the increase in the mobile phone subscriptions. This results in the rise of telecom fraud, which occurs whenever a fraudster performs deceptive methods to get the telephonic services free of charge or at a reduced rate. This problem leads to the loss of subscriber's faith in the service provider company as well as the revenue losses for the organization. According to a study [1] done by Financial Fraud Action United Kingdom (FFA UK), £23.9 million was lost in 2014 in the UK due to various fraudulent activities, which is three times more than the

S. Subudhi · S. Panigrahi (✉)
Department of Computer Science & IT, Veer Surendra Sai University of Technology,
Burla 768018, India
e-mail: spanigrahi_cse@vssut.ac.in

S. Subudhi
e-mail: sharmilasubudhi1@gmail.com

H.S. Behera and D.P. Mohapatra (eds.), *Computational Intelligence in Data Mining*, Advances in Intelligent Systems and Computing 556,
DOI 10.1007/978-981-10-3874-7_60

previous year. The figure shown in the study reflects the growing trend of losses resulting due to rise in fraudulent activities in the telecom industry. Therefore, there is a need to address the mobile phone fraud problem in a quick manner for minimizing the financial losses.

The most common type of telecom fraud is known as the superimposed fraud, which can only be detected by the analysis of a genuine user's account for the presence of any kind of fraudulent activities made by the fraudster by exploiting the genuine account. This type of fraud can remain undetected for a long time as the presence of fraudulent activities is comparatively small in the overall call volume [2]. In this work, we aim at detecting the superimposed mobile phone fraud by applying possibilistic fuzzy c-means (PFCM) clustering on the subscriber's call detail records (CDRs). We have demonstrated the effectiveness of our proposed system by performing extensive experiments on reality mining dataset [3]. To the best of our knowledge, this is the first ever attempt to develop a mobile phone FDS by using PFCM clustering technique.

The rest of the paper is organized as follows: Section 2 discusses the previous work done in mobile phone fraud detection. Section 3 focuses on the fundamental concept of PFCM clustering. The next section elaborates the proposed FDS along with its working methodology. In Sect. 5, we have discussed the results obtained from the experimental analysis. Finally, in Sect. 6, we conclude the paper with some future enhancements of the proposed model.

2 Related Work

In this section, some published works have been reviewed that are relevant to the mobile phone fraud detection problem. In paper [4], the authors have suggested the usage of Dempster–Shafer theory and Bayesian inferencing for information fusion from various sources for the detection of fraudulent activities. The authors of [5] present the application of feed forward neural network and hierarchical agglomerative clustering for fraud detection with five different user profiles for each user. They have applied each technique independently on those user profiles for discriminating illegitimate calls from the legitimate ones by visualizing different aspects of the model. The work in [6] proposes an FDS for the detection of malicious activities present in a subscriber account by data visualization using self-organizing map (SOM).

Another recent work [7] suggests the building of five different user profiles from the features selected by applying four different feature selection algorithms. A genetic programming (GP)-based classifier is then used for the detection of fraudulent patterns. The usefulness of K-means clustering and hierarchical agglomerative clustering algorithms has been presented in [8] for fraud detection. The discrimination of fraudulent signatures from the genuine ones are done by applying these two methods independently on five different user profiles built from the CDRs of each user.

Although several methodologies have been suggested for developing an efficient mobile phone FDS, one of the major issues in the above-mentioned systems is the limited applicability of various hard clustering methods for solving such type of real-world problem in which there is no crisp boundary for segregating the normal user profile and intrusive patterns. Moreover, an individual data point may belong to more than one cluster with different membership values. For improving the accuracy of fraud detection, we have therefore applied the possibilistic fuzzy c-means clustering algorithm in the current work. Besides, this method is superior to two other fuzzy clustering algorithms, namely fuzzy c-means (FCM) and possibilistic c-means (PCM) as it solves the outlier sensitivity problem of FCM and the overlapped cluster issue of PCM.

3 Background Study

In this section, we briefly describe the working principle of PFCM for demonstrating the training and fraud detection methodologies of our proposed system.

3.1 *Possibilistic Fuzzy C-Means Clustering*

Possibilistic fuzzy c-means (PFCM) [9] clustering is a hybrid of two most widely used fuzzy clustering algorithms, namely FCM [10] and PCM [11]. PFCM overcomes the inefficiency of handling noisy instances of FCM and the coincident cluster problem of PCM simultaneously. PFCM takes unlabeled instances of a dataset and attempts to form clusters by finding the most appropriate point as centroid in each cluster. A membership value and typicality value is then assigned to every point in the clusters. This is attained by minimizing the *objective function* as stated below:

$$\begin{aligned} MinJ_{m,\eta}(U,T,V;D) = & \sum_{j=1}^{n}\sum_{i=1}^{c}(a_{ij}^{m}+bt_{ij}^{\eta})\left\|d_j - v_i\right\|_A^2 \\ & + \sum_{i=1}^{c}\gamma_i\sum_{j=1}^{n}\left(1-t_{ij}\right)^{\eta} \end{aligned} \quad (1)$$

subject to constraints $\sum_{i=1}^{c} u_{ij} = 1\forall j$ and $0 \leq u_{ij}, t_{ij} \leq 1$ where $J_{m,\eta}$ is the objective function, m is the fuzzifier weighting exponent, and η is the scale parameter. $D = \{d_1, d_2, \ldots, d_n\}$ is the dataset with n points on which PFCM is to be performed, $U = [u_{ij}]$ is the membership matrix, $T = [t_{ij}]$ is the typicality matrix, $V = \{v_1, v_2, \ldots, v_c\}$ is a matrix of c cluster centers, and $\left\|d_j - v_i\right\|_A$ is the inner product norm used to compute the distance between cluster center v_i and the data point d_j, $\gamma_i > 0$ is a user defined constant value and $a > 0$ is the significance of the membership value, $b > 0$ is the significance of typicality value, $m > 1$ and $\eta > 1$.

PFCM exhibits the FCM properties when $b = 0$ and displays PCM characteristics when $a = 0$. The clustering output becomes more favorable towards PCM as the value of b increases with respect to a and vice versa. For PFCM, a larger value of b is required to be considered as compared to a in order to reduce the effect of outlier instances. Likewise, the effects of noisy points can be reduced for a higher value of m than η. However, a very large value of m can cause the clustering model to be more receptive toward PCM. On giving a dataset as input to PFCM, it produces three different outputs—fuzzy membership matrix (U), typicality matrix (T), and a set of cluster centers (V) computed by using Eq. (1).

4 Proposed Approach

The proposed mobile phone FDS monitors the calling activities of the subscribers by analyzing their CDRs and identifies any fraudulent patterns by applying the PFCM clustering technique. The flow of events in our FDS is partitioned into two phases—training phase and fraud detection phase.

4.1 Training Phase

The training phase deals with the construction of behavioral profile of each user. We have considered the following relevant features for the representation of CDR of a user:

$$\langle user\_id, timedt, dur, type\_call\rangle$$

The feature *user_id* is used to uniquely identify each user by taking the anonymous interpretation of the IMEI (International Mobile Equipment Identity) number of the user's mobile device. The *timedt* denotes the date (ddmmyyyy) and time (hh:min:sec in 24-h format) of a call when it is made. Similarly, *dur* signifies the call duration measured in seconds and *type_call* refers to the type of calls made, which has been mapped to numerals as: 0 for local calls, 1 for national calls and 2 for international calls in our approach. For instance, suppose <7, *29042004171119, 50, 1*> represents a CDR of a user. This example indicates a call record having *user_id* as 7, *timedt* is 29-04-2004 and 17:11:19, *dur* is 50 s and *type_call* = 1 (national).

Initially, we perform normalization on the CDRs by converting all points in the range of [0, 1], since the high-valued attribute fields can cause bias while clustering. The CDRs of a subscriber are partitioned into training and testing sets. Once the segmentation of the dataset is complete, the parameter setting of PFCM is carried out by conducting experiments. The user behavioral profiles are then obtained by employing PFCM clustering technique on the training dataset based on the attributes: *dur* and *type_call*.

4.2 Fraud Detection Phase

After the profile building of a user is successfully completed, the testing set is used in the clustering model for the detection of any kind of fraudulent patterns present in the CDRs. This is accomplished by initially measuring the Euclidean distance (*d*) of the incoming call record with each cluster centroid and finally comparing *d* with a preset threshold value (*th*). The threshold value has been determined through rigorous experimentation as discussed in Table 3 of Sect. 5. If the distance value is higher than or equal to the threshold, then the call is marked as a fraudulent one. On the other hand, if the distance is smaller than the threshold value, then the call is identified as genuine. Upon detecting any illegitimate activities, the service provider company can obtain confirmation regarding the call from the respective subscriber.

5 Experimental Results and Discussions

The proposed FDS has been implemented in MATLAB 8.3 on a 2.40 GHz i5-4210U CPU system. The usefulness of our FDS has been presented by testing with reality mining dataset [3]. Initially, we have performed tests for the determination of optimal parameter values needed for PFCM clustering. Once the required parameters are obtained, we then conduct the fraud detection experiments.

The reality mining dataset consists of call records, messaging records, and much more information of 106 users gathered over a 9-month period from September 2004 to April 2005. We have used the following standard performance metrics—true positive rate (TPR), false positive rate (FPR), accuracy, precision, and F-score to measure the effectiveness of our proposed system. *TPR* denotes the ratio of truly positive samples that are correctly classified by the classifier. *FPR* measures the fraction of rejected genuine samples that incorrectly identify as fraudulent by the classifier. *Accuracy* estimates the correctness of a classifier. *Precision* can be depicted as the proportion of correct classification made by the system, and *F-score* refers to the harmonic mean of precision and TPR.

5.1 Determination of PFCM Parameters

In this section, we discuss the estimation of the correct combination of the clustering parameters required for the working of PFCM. Initially, we perform a set of

Table 1 Determination of optimal number of clusters

c	1	2	3	4
PC	1	*0.9824*	0.9456	0.9346

experiments for finding out the required number of clusters. We have considered partition coefficient (PC) index [12] for measuring the clustering validity as shown in Table 1.

$$PC = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{2} \tag{2}$$

where PC denotes the average relative quantity of membership sharing done between the fuzzy subset pairs in the $U = \left[u_{ij}\right]$ matrix, c is the number of clusters, and n refers to the data points on which clustering is to be performed. The optimal cluster number (c^{+}) is chosen as follows:

$$c^{+} = max_{2 \le c \le n-1} PC \tag{3}$$

which has been presented in italics for better visualization in Table 1. The PC values are computed using Eq. (2), which produces the maximum value, i.e., PC = 0.9824 at $c^{+} = 2$ after satisfying with Eq. (3). Hence, we have chosen the optimal number of clusters $\mathrm{c}^{+} = 2$.

After the correct number of clusters is determined, we then find the optimal combination of the other four PFCM parameters—significance of the membership value (a), significance of typicality value (b), weighting exponent (m), and scale parameter (η). The clustering output of PFCM is presented in Table 2 by taking different combinations of parameters with $c^{+} = 2$ along with a Fuzziness Performance Index (FPI) value. The FPI [13] can be defined as a measurement of the degree to which different classes share membership values. The optimal partition of fuzzy clustering can be found by minimizing the FPI value as this implies that the cluster elements have minimum overlapping between themselves. The FPI value can be calculated as follows:

$$FPI = 1 - (c{*}PC - 1)/(c - 1) \tag{4}$$

where c is the number of clusters and PC is the partition coefficient index. From Table 2, it is quite clear from the cluster centroids $\{v_1, v_2\}$ that except at run 5 and run 6, all other runs produce overlapped clusters. The FPI values are calculated by using Eq. (4). However, the FPI value of run 6 is lesser than the FPI value of run 5. Therefore, we chose the parameter values of a, b, $m,$ and η of Run 6 as an optimal combination of PFCM parameters, which has been italicized in Table 2 for better visualization.

The effectiveness of our proposed system also depends on the threshold value (th). The variations in TPR, FPR, accuracy, precision, and F-score over different threshold values are depicted in Table 3. It is clear from Table 3 that for

Table 2 Results produced by PFCM with different parameter values

Run	a	b	m	η	v_1	v_2	FPI
1	1	1	2	2	0.0082	0.0082	−1.878e+03
					0.5286	0.5286	
2	1	3	2	2	0.0090	0.0090	−1.878e+03
					0.5259	0.5259	
3	1	6	2	2	0.0098	0.0098	−1.878e+03
					0.5190	0.5190	
4	1	7	2	2	0.0099	0.0099	−1.878e+03
					0.5168	0.5168	
5	1	1	5	1.5	0.0097	0.0036	−3.2893 e+03
					0.4437	0.8428	
6	*1*	*1*	*7*	*1.5*	*0.0097*	*0.0018*	*−3.4189 e+03*
					0.4175	*0.9509*	
7	1	5	5	1.5	0.0083	0.0083	−1.878e+03
					0.5065	0.5065	
8	1	5	5	10	0.0087	0.0087	−1.878e+03
					0.5096	0.5096	
9	1	1	2	7	0.0079	0.0079	−1.878e+03
					0.5282	0.5282	
10	1	4	3	2	0.0094	0.0094	−1.878e+03
					05123	0.5123	

Table 3 Variation in different performance metrics over different threshold values

Threshold (*th*)	TPR (in %)	FPR (in %)	Accuracy (in %)	Precision (in %)	F-score (in %)
0.001	90.15	9.83	90.16	91.50	90.82
0.003	*95.07*	*9.25*	*93.09*	*92.34*	*93.69*
0.005	93.30	10.44	91.49	90.50	91.88
0.007	90.50	10.23	90.16	90.95	90.73
0.009	91.00	11.36	89.89	90.10	90.55

th = 0.003, our proposed system exhibits maximum TPR = 95.07% and minimum FPR = 9.25%. Hence, we choose the *th* = 0.003 for efficient fraud detection.

Table 4 presents a comparative performance analysis of various clustering methods on different performance metrics by taking the cluster number $c = 2$. It can be clearly seen that PFCM outperforms other clustering techniques by yielding better performance in terms of all performance metrics while keeping FPR = 9.25% at the lowest level. This selection is essential as the failure to detect a fraud causes direct loss to the service provider while the actions required to handle the false alarms also tend to be costly.

Table 4 Performance analysis of various clustering techniques

Clustering method	TPR (in %)	FPR (in %)	Accuracy (in %)	Precision (in %)	F-score (in %)
PFCM	*95.07*	*9.25*	*93.09*	*92.34*	*93.69*
FCM	75.68	12.34	80.59	89.94	82.15
PCM	84.68	11.69	86.17	91.26	87.85
K-means	69.51	15.03	75.80	87.08	77.31

6 Conclusions

In this work, a novel mobile phone fraud detection system has been suggested by employing PFCM clustering technique. The fraud detection procedure is segmented into two phases—training phase and fraud detection phase. For measuring the efficiency of our system, the reality mining dataset has been used. PFCM clustering algorithm is used for behavioral profile construction of mobile phone subscribers as well as for identification of any intrusive signatures present in their profiles. The experimental results show the ability of PFCM in detecting fraudulent activities of various users. Based upon the outcomes, it can be concluded that by using PFCM clustering technique, this kind of real-world problematic scenario can be addressed effectively.

Acknowledgements The authors are highly grateful to the Department of Computer Science & Engineering and IT, Veer Surendra Sai University of Technology, Burla, Sambalpur, Odisha, India, for providing the required amenities and support for making this successful investigation.

References

1. Kosmides, M,: Telephone fraud on rise in UK, study finds (2014) http://www.counterfraud.com/fraud-types-n-z/telecoms-fraud/telephone-fraud-on-rise-inuk-study-finds–1.htm, accessed: 30 January, 2016.
2. Cox, Kenneth C., et al.: Brief application description; visual data mining: Recognizing telephone calling fraud. Data Mining and Knowledge Discovery1.2 (1997) 225–231.
3. Eagle, N., Pentland, A.S.: Reality mining: sensing complex social systems. Personal and ubiquitous computing 10.4 (2006) 255–268.
4. Panigrahi, S., et al.: Use of dempster-shafer theory and Bayesian inferencing for fraud detection in mobile communication networks. Australasian Conference on Information Security and Privacy. Springer Berlin Heidelberg, (2007).
5. Hilas, C.S., Paris A.M..: An application of supervised and unsupervised learning approaches to telecommunications fraud detection. Knowledge-Based Systems 21.7 (2008) 721–726.
6. Olszewski, D.: Fraud detection using self-organizing map visualizing the user profiles." Knowledge-Based Systems 70 (2014) 324–334.
7. Hilas, C.S., et al.: A genetic programming approach to telecommunications fraud detection and classification. (2014).

8. Hilas, C.S., Paris A.M., Ioannis T.R.: Clustering of Telecommunications User Profiles for Fraud Detection and Security Enhancement in Large Corporate Networks: A case Study. Applied Mathematics & Information Sciences 9.4 (2015) 1709.
9. Pal, N.R., et al.: A possibilistic fuzzy c-means clustering algorithm. IEEE transactions on fuzzy systems 13.4 (2005) 517–530.
10. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences 10.2–3 (1984) 191–203.
11. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE transactions on fuzzy systems 1.2 (1993) 98–110.
12. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy sets and systems 158.19 (2007) 2095–2117.
13. Odeh, I. O. A., Chittleborough, D. J., McBratney, A. B.: Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of America Journal 56.2 (1992) 505–516.