International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Identifying Fraudsters and Fraudulent Strategies in Mobile Social Network

J. Jyotsna [1], A. Shruthi[2], G. Bhanu Prakash Reddy[3], V. Srikanth[4]

[1]Assistant Professor, Department of Information Technology, J.B. Institute of Engineering and Technology, Hyderabad
[2,3,4]Student, Department of Information Technology, J.B. Institute of Engineering and Technology, Hyderabad

Abstract
Moderncommunication technologies have developed quickly, especially communications through (mobile) phones, which has greatly aided in the sharing of information and social connections between people. Nonetheless, the rise of telemarketing scams has the potential to seriously deplete communal and private wealth, slowing down or harming the economy. With an emphasis on exposing the "precise fraud" phenomena and the techniques employed by fraudsters to precisely choose targets, we propose to identify telemarketing scams in this study. We utilise a one-month comprehensive dataset of telecommunication information from Shanghai, which includes 698 million call logs and 54 million customers, to explore this issue. During our research, we have discovered that user information may have been substantially compromised, and that fraudsters prefer to target users who are younger and more active on mobile networks. To further separate fraudsters from non-fraudsters, we provide a unique semi-supervised learning approach. Our technique beats various cutting-edge algorithms in terms of accuracy of identifying fraudsters, according to experimental findings on real-world data. We think that our research may help governments and mobile service providers make better policy decisions.

Keywords: Precise fraud, Semi-Supervised Machine Learning

1. Introduction

With the recent advancements in worldwide communication technologies, fraudulent activities are growing quickly. Frauds cause terrible suffering to millions of people. For instance, phone fraud has been recognised as a serious issue in China. According to estimates from Qihoo1 and Tencent2, there were over 500 million phone scams in 2016, resulting in losses of over 16.4 billion USD. Less than 3% of these cases are eventually settled. A college professor in Beijing was alleged to have lost $2.67 million on August 29th, 2016, to a phone scammer posing as a judge. Apart from the financial toll that phone fraud has had on individuals, the results have been considerably more devastating, even lethal. Most of the current research on fraud detection designs trials using fictitious data or small-scale real-world data. In this study, we examine a sizable mobile social network in the real world, which spans 30 days from September 1 to September 30 and includes an entire collection of call records from Shanghai. The anonymous phone numbers and the beginning and finish times of each call are noted in the call log. We also get crowd-sourced annotations on scammers. Mobile telecom fraud describes unauthorised ac-

cess to a mobile operator's network and the use of its services for illicit purposes at the expense of the network's operators or its users. During our research, we have discovered that user information may have been substantially compromised and that fraudsters prefer targeting users who are younger and more active on mobile networks. To further separate fraudsters from non-fraudsters, we provide a unique semi-supervised learning approach. Experimental findings on a real-world data reveal that our technique surpasses various state-of-the-art algorithms in accuracy of detecting fraudsters. The system should be able to collect test data results and offer the user a final output based on the context of the test data. We create the unique factor graph-based model FFD to identify fraudsters based on our findings. Our approach explicitly considers the structural knowledge and target choice of fraudsters. To address the label sparsity concern, we also provide a semi-supervised learning architecture that makes use of both known and unknown labels. According to our tests, our model outperforms various cutting-edge techniques, with an improvement on F1 of 0.278.

It is important to emphasise the following aspects of our contributions:

o We describe how fraudsters and non-fraudsters act differently in mobile networks based on real phone-communication data.

o We research the "exact fraudulent approach" and issue a call to action for everyone to prioritise the security of private data.

o To differentiate fraudsters from other users in a certain mobile network, we offer a unique architecture.

o We test the performance of our model on a sizable mobile network in the real world.

2. Literature Survey

How can we identify online fraudsters that use big IP pools (with as many as 852,992 dedicated IPs) to disguise fraudulent assaults while manipulating geolocations, internet service providers, and IP addresses? According to research from December 2016, online fraud has become a severe issue due to the enormous potential reward it gives to fraudsters, which may be as much as $5 million from 300 million phoney "views" every day. Let us say a fraudster services a customer who wants to purchase 200 ratings or clicks for each of b items. The fraudster may have several accounts or IPs. Like with other systems, we assume that each account can only rate a product once. Given that there are 200 alternative ratings for each product, the density of the constructed fraudulent block is (200b)/(ab) = 200/a. The fraudster can therefore provide as many items as necessary while maintaining a low density if they have access to enough user accounts or IPs. Most current fraud detection techniques have a tough task as a result of this [1]. With the advancement of contemporary technology and international communication, fraud is fast rising. Millions of individuals therefore endure terrible suffering as a result of these fraudulent actions. Many anti-fraud mechanisms have been put out to shield individuals from the harm caused by fraudulent actions, however such systems for communications are still in their infancy. Almost all the anti-fraud techniques used in telecommunications rely on crowdsourced annotations. In other words, before a phone number is blacklisted as a fake, fraudsters can still call innocent individuals using it. It's feasible for fraudsters to continue making phone calls using a phone number for days, weeks, or even months because not everyone is ready to annotate it. This occurrence is what we refer to as the time lag in the fraud detection issue. Dealing with the issue of the time lag in fraud detection for the identification of fraudulent phone calls has therefore become a crucial subject to be investigated [2]. This essay's goal

is to investigate several facets of telecom fraud detection and prevention. In this paper, many types of telecommunication fraud are reviewed, along with the obstacles that prevent their detection and some suggested strategies to solve them. In addition to scalability and effectiveness, the fraud-detection job displays technical issues such as skewed training data distributions and nonuniform cost per error, neither of which have received much attention from the knowledge-discovery and data mining communities. In this article, we review and assess several approaches that simultaneously solve these three major problems.

3. System Architecture

System architecture diagram is an abstract representation of the component architecture of the system. It gives a concise explanation of the system's component architecture to aid with component-component connections.

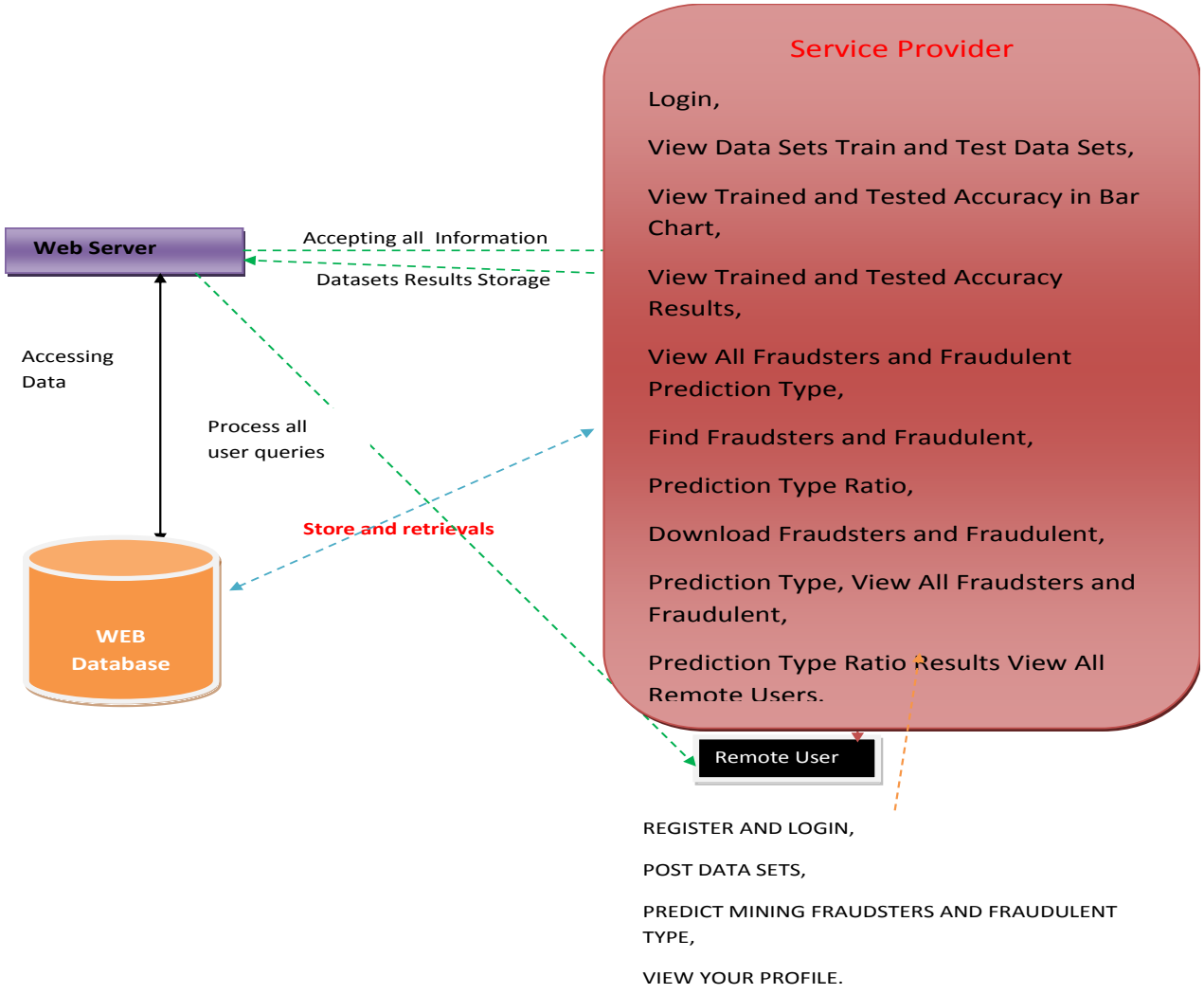

Fig. 1 System Architecture for Proposed System

The system architecture diagram depicts the system architecture visually. It depicts the links between the system's many components and identifies what functions each component performs. The general system representation depicts the system's primary operations as well as the links between the different system components

4. Algorithm

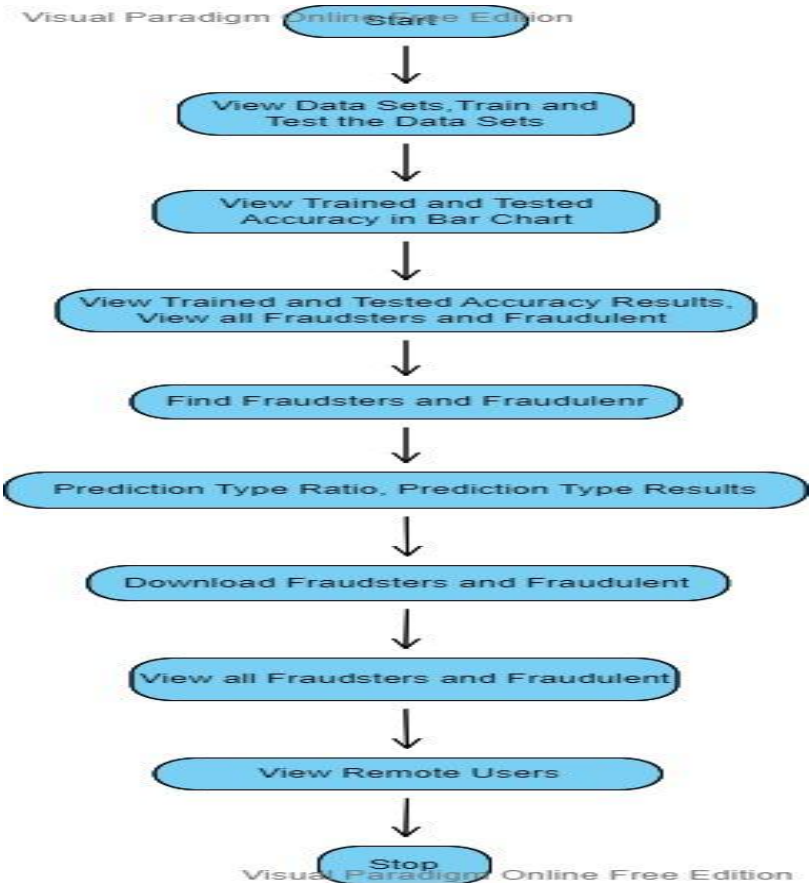

Fig. 2 Flow Diagram of the Proposed system

A. SVM Algorithm:

"A support vector machine (SVM) is a supervised machine learning model that solves two-group classification problems using classification techniques. They can categorise fresh text after providing an SVM model with sets of labelled training data for each category.

Assume we have two tags, red and blue, and two features, x and y, in our data. We want a classifier that can determine if a pair of (x,y) coordinates is red or blue. On a plane, we plot our already labelled training data:

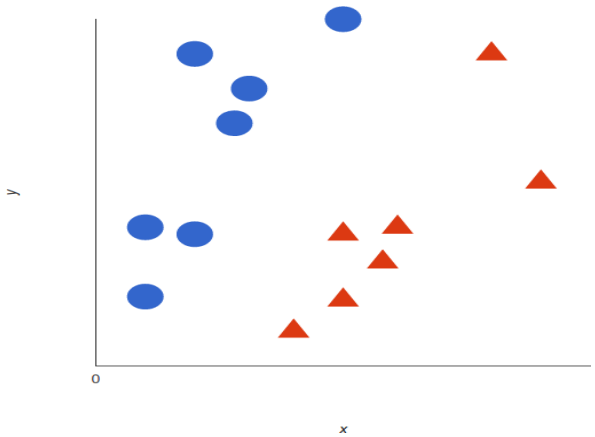

Fig.3 labelled data

A support vector machine takes these data points andproduces the hyperplane (which is basically a line in two dimensions) that best separates the tags. This line represents the decision boundary: everything on one side will be classified as blue, and anything on the otherwill beclassifiedasred.

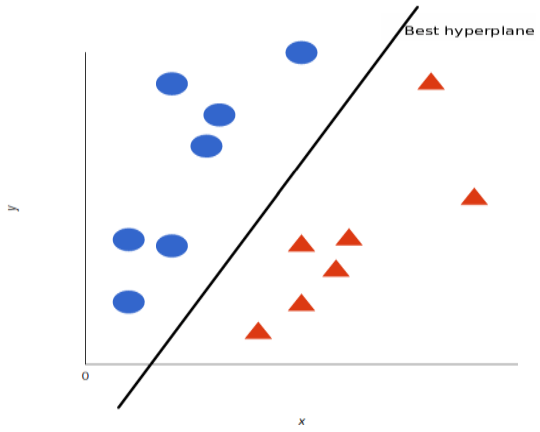

Fig.4.Classification ofdata

The hyperplane (remember, it's a line in this case) with the greatest distance to the nearest element of each tag [10].

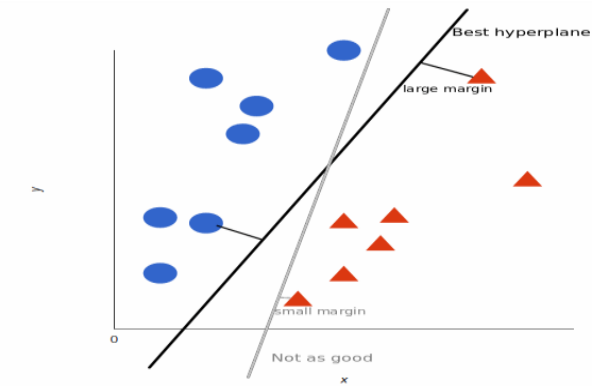

Fig.5AnExample ofWorkingoftheAlgorithm

Decision tree classifiers:

Decision tree classifiers have a wide range of effective applications. Their capacity to extract descriptive decision-making information from the provided data is their key strength. Training sets can be used to create a decision tree. The process for such creation is as follows, and it is based on a collection of objects (S), each of which belongs to one of the classes C1, C2…., Ck:

Step 1: The decision tree for S has a leaf labelled with this class if all the objects in S are members of the same class, such as Ci.

Step 2: If not, let "T" be any test with potential results O1, O2..., On. The test divides S into subsets S1, S2..., Sn where each object in Si has result Oi for T since each object in S has one outcome for T. T becomes the decision tree's root, and we create a subsidiary decision tree for each result Oi by repeating the same steps on the set Si.

5. Conclusion and Future Scope

In this article, we investigate the issue of fraudsters and fraudulent tactics in a sizable mobile network. We discover that fraudsters and non-fraudsters behave differently while speaking with others after studying a one-month comprehensive dataset of telecommunication information in Shanghai with 698 million call records between 54 million individuals. Also, while picking targets, scammers have preferences for individuals' ages and phone usage patterns. We next suggest a brand-new semi-supervised model to identify fraudsters from non-fraudsters, which is motivated by our exploratory investigation. Experimental findings show that our model significantly outperforms several state-of-the-art baseline techniques. In terms of future work, it's intriguing to consider how to identify fraud groups, which, as opposed to a single fraudster, comprise of fraudsters with a variety of roles and responsibilities. This allows for the disclosure of several fraud organisations' patterns of coordination. Moreover, one may expand on our findings by considering user geography and researching offline fraudster activities including their city-hopping habits. The available data puts a limit on what we can accomplish. Although Shanghai is a significant worldwide metropolis and China Telecom is a significant service provider, the selection bias in our data may restrict the generalizability of our study.

6. References

[1] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, “FRAUDAR: Bounding graph fraud in the face of camouflage,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 895–904.

[2] V. S. Tseng, J. Ying, C. Huang, Y. Kao, and K. Chen, “FrauDetector: A graph-mining-based framework for fraudulent phone call detection,” in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 2157–2166.

[3] J. Xu, A. H. Sung, and Q. Liu, “Behaviour mining for fraud detection,” J. Res. Practice Inf. Technol., vol. 39, pp. 3–18, 2007.

[4] M. I. M. Yusoff, I. Mohamed, and M. R. A. Bakar, “Fraud detection in telecommunication industry using Gaussian mixed model,” in Proc. Int. Conf. Res. Innovation Inf. Syst., 2013, pp. 27–32.

[5] P. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, “Distributed data mining in credit card fraud detection,” IEEE Intell. Syst. Appl., vol. 14, no. 6, pp. 67–74, Nov./Dec. 1999.

[6] T. Ormerod, N. Morley, L. Ball, C. Langley, and C. Spenser, “Using ethnography to design a mass detection tool (MDT) for the early discovery of insurance fraud,” in Proc. Extended Abstracts Human Factors Comput. Syst., 2003, pp. 650–651.

[7] Y. Yang, C. Tan, Z. Liu, F. Wu, and Y. Zhuang, “Urban dreams of migrants: A case study of migrant integration in Shanghai,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 507–514.

[8] Y. Yang, Z. Liu, C. Tan, F. Wu, Y. Zhuang, and Y. Li, “To stay or to leave: Churn prediction for urban migrants in the initial period,” in Proc. 27th World Wide Web Conf., 2018, pp. 967–976.

[9] Y. Yang, J. Tang, and J. Li, “Learning to infer competitive relationships in heterogeneous networks,” ACM Trans. Knowl. Discovery Data, vol. 12, 2018, Art. no. 12.

[10] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, “Inferring user demographics and social strategies in mobile social networks,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 15–24.

[11] Y. Yang, J. Tang, J. Keomany, Y. Zhao, J. Li, Y. Ding, T. Li, and L. Wang, "Mining competitive relationships by learning across heterogeneous networks," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1432–1441.
[12] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," Proc. Nat. Academy Sci. United States America, vol. 106, no. 51, pp. 21 544–21 549, 2009.
[13] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Trans. Inf. Theory, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[14] J. M. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," Unpublished Manuscript, 1971