Anomaly Detection from Call Data Records

Nithi and Lipika Dey

TCS Innnovation Labs, Delhi
{nithi.1,lipika.dey}@tcs.com

Abstract. In this paper, we propose an efficient algorithm for anomaly detection from call data records. Anomalous users are detected based on fuzzy attribute values derived from their communication patterns. A clustering based algorithm is proposed to generate explanations to assist human analysts in validating the results.

Keywords: Anomaly detection, Fuzzy logic, DBSCAN Algorithm, Principal Component Analysis.

1 Introduction

Call Data Records (CDR) define a special kind of social network where nodes represent phone numbers and an edge represents a call between two numbers. Social networks derived from call data records model relationships among a group of mobile users. Investigative analysis has come to heavily rely on Call Data Record (CDR) analysis since these can provide major cues about temporally and spatially allied events as well as people involved.

Anomaly detection from call data records employs data mining techniques to identify abnormal behavioral patterns. As of today, major emphasis of CDR analysis has been towards visual analytics [1, 2]. The major drawback of such systems is the over-emphasis on the analyst's capability to identify the regions of interest, particularly if the volume of data to be considered is prohibitively large.

In this paper we propose an automated anomaly detection mechanism that assigns anomaly score to nodes in a CDR network. It has been shown that rather than using the features of a single subscriber and his contacts, better results can be obtained by using attributes of links at greater depths. Efficiency is ensured through suitable feature transformation techniques.

2 Anomaly Detection in Social Networks – A Review

Anomaly detection is a mature area of research [3] and has been successfully applied to various areas like network intrusion detection, credit-card fraud detection, email based network analysis etc. They work on the premise that normal behavior is more pre-dominant than abnormal behavior and will be exhibited by majority of the network entities. Commonly used techniques for network analysis involves computation of Centrality measures [4]. [5] had proposed the use of indirect connections to detect

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 237–242, 2009.

anomalous subscribers from call data records. [6] had applied the algorithm to VAST dataset [9] to discover suspicious elements.

3 Feature Identification for Anomaly Detection

The properties typically used for Call Data Record analysis are those that capture individual user behavioral properties like number of calls made or received, number of contacts, towers used for making calls etc. Users can also be associated with statistical properties like average number of calls made per day, average number of contacts, etc. Interaction between any two users is described by the nature of calls between them. In our approach, behavior of an ID is derived from his direct and indirect interactions.

Call details from a service provider contain information about calls made or received by its subscribers. Each call is represented by a caller ID, a receiver ID, call duration, time of call initiation, tower used, call type (VOICE or SMS) and some details about the handset used. Each interaction is either of type OUTGOING or INCOMING for a given user. We represent the behavior of each user by his calling pattern and also his interaction patterns with other users. Interaction pattern takes into account the frequencies of different types of calls associated to the user.

The most important attribute of a call is its duration. Rather than using exact call duration times we characterize a call as “long” or “short” with a certain degree of fuzziness. They provide a robust characterization which is more tolerant to noise.

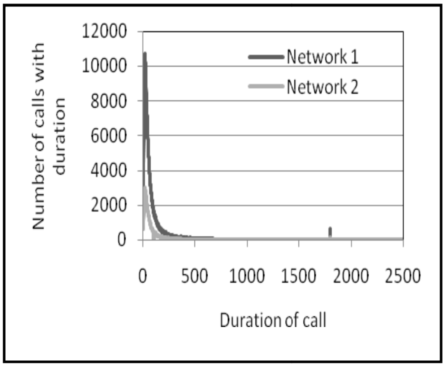

Fig. 1. Distribution of the duration of calls for two networks

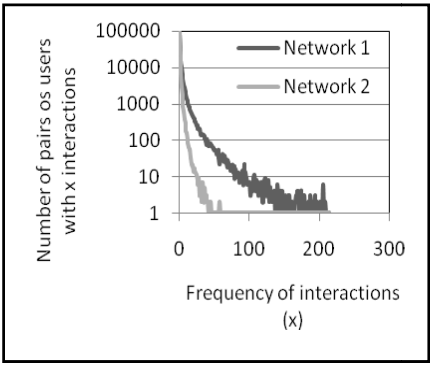

Fig. 2. Distribution of the number of pair-wise interactions

The relevant fuzzy membership functions were derived after analyzing various sets of call data records for different mobile networks. Figure 1 shows graphs for two networks - network 1 with 30866 users, 77691 calls over 12 days and network 2 with 31494 users, 207704 calls over 2 days. It shows that short duration calls are much more frequent than long duration calls. The exact cut-off value for identifying a call as long or short however can vary for different networks since they operate with different pulse rates. The proposed fuzzy membership function is designed to take this variation into account. It is as follows.

A long duration call is one which is longer than 95% or more of the other calls and a short duration call is one whose duration is less than the duration of 75% of the calls. A call of duration d is characterized as follows:

$\mu_S(d) = 1, \mu_L(d) = 0$ if the d < 75 percentile of duration of calls

$\mu_S(d) = 0, \mu_L(d) = 1$ if the d > 95 percentile of duration of calls

$$\mu_S(d) = \frac{0.95 - x}{0.95 - 0.75}, \mu_L(d) = \frac{x - 0.75}{0.95 - 0.75} \text{ where x denotes the percentile of d} \quad (1)$$

otherwise

where S denotes short, L denotes long.

While a call can be characterized by its duration, interactions between two users can be defined by the frequency of different types of calls between them. Frequency can be characterized as "high" or "low". Figure 2 shows that pair of users that interact infrequently is exceptionally more than pairs with frequent interactions.

Depending on the frequency f of interaction between two users, the interaction type is characterized as "high" or "low" using the following fuzzy functions:

$\mu_{low}(f) = 1, \mu_{high}(f) = 0$ if the f < 95 percentile of frequency of calls

$\mu_{low}(f) = 0, \mu_{high}(f) = 1$ if the f > 99 percentile of frequency of calls

$$\mu_{low}(f) = \frac{0.99 - x}{0.99 - 0.95}, \mu_{high}(f) = \frac{x - 0.95}{0.99 - 0.95} \quad (2)$$

where x denotes the percentile of f, otherwise

We can now combine call characteristics and interaction characteristics to have a more complete characterization of users. Two users can be connected through different types of calls with different frequencies. For example, two users A and B may frequently communicate through "short" calls, but rarely through "long" calls. It may be further observed that most of the "short" calls are initiated by B while the "long" calls are initiated equally by both. Consequently, the interaction pattern of A with B can be characterized as the following vector of fuzzy-membership values:

$<\mu_{LOW}(OS), \mu_{HIGH}(OS), \mu_{LOW}(OL), \mu_{HIGH}(OL), \mu_{LOW}(IS), \mu_{HIGH}(IS), \mu_{LOW}(IL), \mu_{HIGH}(IL)>$, where the symbol O is used for an OUTGOING call and I for an INCOMING call, S stands for short-duration call, L for long-duration call. $\mu_{LOW}(OS)$ is computed using the weighted average of fuzzy memberships of all OUTGOING calls to the class "short" or S. All other memberships are computed in a similar fashion.

It may be essential to consider a user's indirect connections also. In the current framework, if a user is connected to another through a chain of contacts, then the indirect interaction pattern between the two users is defined by the intermediate chain of direct interaction patterns. Let $LINK_1$ represent the direct interaction vector of A with B, and $LINK_2$ represent the direct interaction vector of B with C, then a vector $LINK_{12}$ that defines the indirect interaction pattern of A with C is computed as a 64 component vector. The $(ij)^{th}$ component of $LINK_{12}$ is computed as follows:

$LINK_{12}(ij) = \min(LINK_1(i), LINK_2(j))$, where i and j vary from 1 to 8.

Paths up to any predefined depth can be considered, though the impact of one node on another decreases with increase in depth. The dimensionality of the feature space also goes up. Weight for a pattern for an ID is the sum of weights of all interactions of that pattern for the ID. The behavior space is then normalized with respect to patterns to equalize the importance of all patterns.

For example, consider the call graph described in Figure 3. It is observed that A interacts with B through short duration calls only. Moreover, the frequency of short calls initiated by A is high but terminated at A is low. Consequently, the interaction pattern of A with B is described by the vector stated above with value 1 for μ_{HIGH}(OS) and μ_{LOW} (IS). Further, B interacts with D by initiating long duration calls with medium frequency. Value for μ_{LOW} (OL) is found to be 0.4 and for μ_{HIGH} (OL) is 0.6 in the vector describing interaction of B with D. The interactions of all other users are described in a similar way. Though A is not directly connected to D, but A is connected to D through B and C. Since A calls B by frequent short duration calls and B initiates medium long duration calls to D, therefore the value for the feature μ_{LOW}(OS) μ_{HIGH} (OL) in the interaction pattern of A and D via B is 0.6. Feature values for A's interaction with D through C can be similarly computed. The final values for such an interaction vector are computed using component-wise summation.

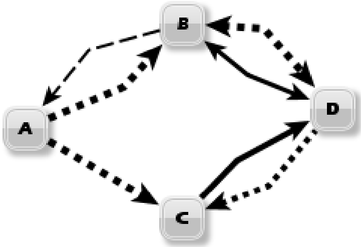

Fig. 3. Call graph. Dotted (solid) lines indicate low (high) duration calls. Frequency of calls is mapped to thickness of the edge.

4 Anomaly Detection

Due to the high dimension of the original feature space derived in the earlier section, we have applied Principal Components Generator to identify the dimensions along which the data set shows maximum variation. We then chose the three principal components only to represent the transformed dataset. An algorithm based on Ramaswamy's distance is designed to detect outliers [8]. This algorithm finds top *n* outliers based on their distance from their kth closest neighbor. This algorithm works as follows:

Step 1: Choose values of n and k.
Step 2: For each data point find its distance from its kth closest neighbor.
Step 3: Arrange the data points in descending order of the distance.
Step 4: Choose top n points as outliers.

5 Explanation of Interaction Patterns of Anomalous Users

Principal components are linear combinations of attributes. They do not retain original feature values. However, for intelligence analysis it is essential that explanation of anomalous behavior be generated in terms of original features. For small data sets one can generate the rules using decision trees, where top *n* anomalies are treated as elements of one class and the rest of the elements of another.

The decision tree is not suitable for generating rules from large datasets. In this case, for each of the top *n* anomalies detected, we find its p closest neighbors in the original feature space. We have set p to 1000. These neighbors are now clustered using the DBSCAN algorithm. Non-anomalous entities are much more in number and they form dense clusters. The features that distinguish the seed anomalous entity from all its neighboring clusters with more than 10 elements are extracted. For each cluster 5th and 95th percentile of the feature values are computed for all features. Each feature value of the seed is then compared with computed percentile of the corresponding features. The features with high variation are then extracted and reported.

6 Experimental Results

Table 1 shows that our system is able to better the results of [6] on VAST 2008 dataset [9]. This set has 400 entities with 10000 interactions. The challenge was to identify 12 members of a gang who conducted suspicious activities. All other results reported in VAST were for human-driven visual analysis [1]. Figure 4 shows that the anomalous entities become easily separable on the transformed space. Figure 5 shows a sample rule characterizing anomalous and non-anomalous entities. A feature characterizing interaction at depth 3 turns out to be most significant. It typically characterizes interaction patterns observed between key gang members and ground level workers.

Table 1. Performance results for various algorithms on VAST2008 dataset

Algorithm	Proposed Algorithm	Kojak[6]
Precision	80%	60%
Recall	80%	60%

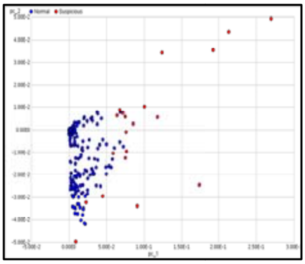

Fig. 4. Distribution of anomalous (RED) and non-anomalous (BLUE) entities (VAST 08

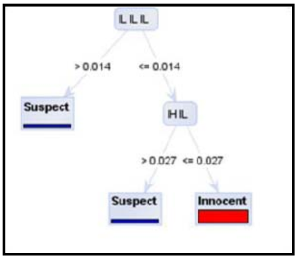

Fig. 5. Rule characterizing anomalous IDs

7 Conclusion

Call data record analysis have gathered momentum due to the role they play in investigative analysis. Most of the research till date has been directed towards visual analytics. We have presented here methods for finding anomalous behavior from large datasets, considering interaction up to any depths. We have presented the viability of using very complex feature spaces in conjunction with PCA to capture complex notions of behavior. The algorithms proposed here have been applied to large real-life data sets and found to scale extremely well.

References

1. Swing, E.: Solving the Cell Phone Calls Challenge with the Prajna Project. In: IEEE Symposium on Visual Analytics Science and Technology, Columbus, Ohio, USA (2008)
2. Payne, J., Solomon, J., Sankar, R., McGrew, B.: Grand Challenge Award: Interactive Visual Analytics Palantir: The Future of Analysis. In: IEEE Symposium on Visual Analytics Science and Technology, Columbus, Ohio, USA (2008)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. To Appear in ACM Computing Surveys (2009)
4. Wasserman, S., Faust, K.: Social Network Analysis: Methods & Applications. Cambridge University Press, Cambridge (1994)
5. Lin, S., Chalupsky, H.: Discovering and explaining abnormal nodes in semantic graphs. IEEE Transactions on Knowledge and Data Engineering 20 (2008)
6. Chalupsky, H.: Using KOJAK Link Discovery Tools to Solve the Cell Phone Calls Mini Challenge. In: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, Portugal (2008)
7. Ross, T.J.: Fuzzy Logic with Engineering Applications. Wiley, Chichester (2004)
8. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient Algorithms for Mining Outliers from Large Data Sets. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Texas, United States (2000)
9. IEEE VAST Challenge (2008), http://www.cs.umd.edu/hcil/VASTchallenge08