International Journal on Advanced Science Engineering Information Technology

Vol.4 (2014) No. 1
ISSN: 2088-5334

Determining an Appropriate Weight attribute in Fraud Call Rate Data Using Case Based Reasoning

Bala Musa Shuaibu

[#] *Computer Engineering and Science Department Yanbu University College, P.O. Box: 31387, 41912, Yanbu Industrial, Saudi Arabia*
E-mail: balam@rcyci.edu.sa

***Abstract*— Fraud cases are significantly causing huge revenue losses in telecommunication companies around the world. Although previous cases are very important data in dealing with fraud patterns, there are variations in the dataset of different fraud case scenarios which in turns need specific detection system without necessarily involving the domain expert directly. This paper investigates the appropriate weight values for attributes using fraud Call Rate Data that is based on Artificial Intelligence technique (Case Based Reasoning) with a meaningful confidence in telecommunication data. The experimental result on the fraud data reports that the weight for all attribute used in this study needs to be set at 0.9 in order to get the best performance of 98.33%.**

***Keywords*— Case Based Reasoning; Fraud; Weight Attribute; Call Rate Data; Telecommunications; Similarity**

I. INTRODUCTION

Telecommunications has brought enormous achievements in terms of provision of distance communication and data transmission through computer network, radio and television. However, telecom companies are faced with huge revenue leakages as a result of fraudsters using the telephone devices in an illegitimate ways to avoid or reduce the charges attached to it. As a result of these activities of fraudsters, some operators are unable to stand the competitive environment coupled with lack of customer's confidence, poor services in terms of congestion.

A report [1] has shown that the average loss in the entire global telecom for the year 2006 has risen to about 12.1% of turnover ($176 billion), compared to 2005 which was 11.6%. Hence fraud was reported as the major area of revenue loss with about 2.9% of the turnover.

Therefore, to mitigate this fraud practices, several approaches that implements artificial intelligence techniques such as using rule based system were used [2], Neural Networks has also been implemented to classify fraud cases [3] and [4]. But giving the inconsistency nature of the fraud patterns in different telecom organization, it calls for continues refinement of methods or approach. The Case Base Reasoning CBR approach gives a framework for case retrieval, reuse, and solution testing and learning. It is rather dependent on the knowledge of previous experience to solve similar cases than the knowledge of the problem itself. And its learning is incremental since the new cases are retained for future purpose [5].

This study attempts to explore the use of Case Based Reasoning in fraud detection system based on the uniqueness of the attribute and their corresponding values. Its performance was measured based on the similarities percentages produced by the prototype. Therefore, to be able to have a meaningful, conclusive and absolute detection, call patterns of a subscriber are observed and a comparison is done between the historical data and the new data to determine certain changes in the pattern of call. The analysis is carried out with sample of usage pattern of CDR (Call Detail Records) from Malaysian Telecommunications.

II. RELEVANT WORK

Initial works on fraud detection concentrate on Rule Base such as [6] that uses the idea of having a threshold based on what they call "acceptable changes" that a customer will make from his previous calling behaviour. Therefore any deviation from such threshold is red flagged and considered fraud. Their approach differs from other approach in the sense that they build a generative model to learn the pattern of customer call over some periods and they formulate the threshold. Also, [7] investigates a fraud system for test bed environment considering two issues: accounting and security. Their approach was typically aimed at finding the process and procedures used by telecom fraud analyst in determining their rules for triggering alarm as fraud in systems. They contended that the rule base techniques are based on absolute or differential rules. The absolute rules are based on simple threshold while the differentials are based on statistic anomalies.

While [8] used a user profiling method that differs from one account to another and where some set of rules are created as a line between fraud cases and non-fraud cases. Therefore, the threshold used to determine fraud cases for one account is different from that used for other accounts.

Other works such as [9] uses neural network to learn classes of fraud and non-fraud, and the uses the Gaussian to model the pattern of subscribers' behavior and in the end applied the Bayesian network to determine what is termed fraud and non-fraud cases.

Similarly, [10] investigated the use of neural network in determining phone pattern usage in order to detect a fraud call. They captured the required data for use in detecting fraud within Call Rate Data (CDR) record. They evaluated fraud based on either intrinsically fraudulent or anomalous where the later needed a neural network approach to detect.

However, [11] combined data mining approach to extract relevant data from a scattered data which is then used as a previous case for solving new problem. They combined technologies of data mining and case base reasoning (CBR) to solving problems. Whereas, [12] also uses the multiple-algorithmic and adaptive CBR technique for fraud classification and filtering of large, noisy data sets to reducing the number of final-line fraud in credit card.

III. Approach

Case Base Reasoning is a method of solving problem by inferring on previous similar cases for re-use. This gives a meaningful confidence and a high accuracy in terms of system measures, the knowledge of expert is not necessary in this regards also. Therefore, this study explores the use of Case Based Reasoning in fraud detection system. Its performance was measured based on the similarity percentages produced by the prototype system.

A. Data Set

The study uses a sample data set from a Malaysian Telecom Operator with more than six thousand records of fraud alarm cases over a period eight month and was pre-processed to select the attributes of interest to develop the system. It is not the intent of this study to outline the pre-processed work on the data. The table 1.0 shows the attributes and its description while figure 1.0 shows the sample data set.

TABLE I
SELECTED ATTRIBUTES OF PRE-PROCESSED DATA

Attribute	Description
Alarm Code	The Code responsible for trigger
Severity	Magnitude of the code
User_group	Subscribers group based on type of calls
User_Confidence	Text
User_Usage	Usage pattern of subscriber
Case_Indicator	Text

	D	E	F	G	H	I
1	ALARM_CODE	SEVERITY	USER_GROUP	USER_CF	USER_USG	IND_CASE
2	038	0	R10	10	180.06667	C
3	033	1	DEF	6	30.16667	C
4	020	1	R10	10	101.15	C
5	029	3	R10	10	123.23336	C
6	038	4	R10	10	187.29998	C
7	021	5	R10	10	105.18	C
8	023	5	DEF	10	210	C
9	023	5	DEF	10	210	C
10	029	6	R10	6.667	127.34999	C
11	023	6	R10	10	106.05	C
12	029	6	R10	10	126.86667	C
13	033	8	R10	6.8	48.43333	C
14	033	9	R10	6.8	49.13333	C
15	037	9	R10	8.571	195.34996	C
16	010	10	DEF	3.333	11	C
17	018	10	R10	10	44	C
18	033	13	DEF	6.4	33.81667	C
19	023	13	R10	10	112.7	C
20	023	13	R10	10	113.4	C
21	009	14	DEF	5.333	205.28335	C

Fig 1. Sample Data Set

B. Case Retrieval Process

The existing best practice cases in the case base is matched with the new cases using a matching component to retrieve a closely or similarity cases. This approach identifies basically similar cases using the similarity measure while assigning weight function to determine Global Similarity values and hence the matching component is computing the similarity between attribute values.

Similarity is based on:

$$\text{Non Numeric sim } (a,b) = 1 \text{ if } a=b \quad (1)$$

$$\text{Sim } (a,b) = 0 \text{ if } a \neq b \quad (2)$$

$$\text{Global similarity } (\%) = 1/\sum w \,[ip \sum wi*sim\,(ai,bi)] *100 \quad (3)$$

IV. Result

This section highlights the result of the prototype and the experiment carried to determine the weight using different cases. The CBR approach obtains or retrieves data from the history cases in the case base that has similarity to the new case by applying the similarity computation with the given weights on the data. It also provided the flexibility for user to adapt or retain new case of fraud based on the expert specification of certain threshold similarity values.

The figure 2 shows the fraud engine retrieval page and figure 3 shows the percentage similarity when all attributes are set to an initial weight of 1.

Fig 2. Fraud Engine Retrieval Page

To study the effect of various weight on similarity performance, several test cases has been design in order to determine the suitable weight for each attribute. Such cases are explained in the summary table 2 and 3.

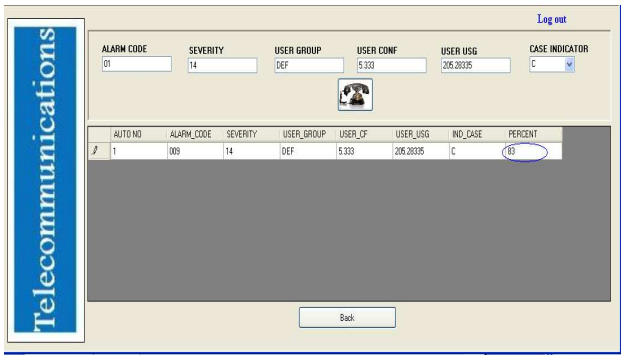

Fig 3. Similarity when all attributes are set to an initial weight of 1

Therefore the percentage similarity retrieved by the engine is 83 and upon further investigation, the weight for the first attribute is then decreased to 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2 and 0.1. The result as shown in table 2 is higher with weight value of 0.9 at a percentage of 98.33.

TABLE II
SIMILARITIES FOR FIRST ATTRIBUTE

Weight of First Attribute	Percentage Similarity
0.9	98.33
0.8	96.67
0.7	95.00
0.6	93.33
0.5	91.67
0.4	90.00
0.3	88.33
0.2	86.67
0.1	85.00

The best weight is fixed for the first attribute and further investigation is carried out for the rest of the attributes with similarities as shown in Table 3.

TABLE III
SIMILARITIES FOR THE REST OF ATTRIBUTE

Weight	Attribute 2 similarity when attribute 1 set to 0.9	Attribute 3 similarity when attribute 1 and 2 set to 0.9
0.9	96.67	95.17
0.8	95.00	93.33
0.7	93.33	91.67
0.6	91.67	90.00
0.5	90.00	88.33
0.4	88.33	86.67
0.3	86.67	85.00
0.2	85.00	83.33
0.1	83.33	81.67

Weight	Attribute 4 similarity when attribute 1,2 and 3 set to 0.9	Attribute 5 similarity when attribute 1,2,3 and 4 set to 0.9	Attribute 6 similarity when attribute 1,2,3,4 and 5 set to 0.9
0.9	94.22	91.67	90.00
0.8	93.33	90.00	88.33
0.7	91.67	88.33	86.67
0.6	90.00	86.67	85.00
0.5	88.33	85.00	83.33
0.4	86.67	83.33	81.67
0.3	85.00	81.67	80.00
0.2	83.33	80.00	78.33
0.1	81.67	78.33	76.67

It is observed that all the attributes perform well when the weight is set to 0.9. This obviously indicates that the weight for all attribute used in this study needs to be set as 0.9 in order to get 98.33% similarity performance.

V. CONCLUSIONS

This study explores the use of Case Based reasoning technique to measure based on the similarity percentage produced by the prototype, the appropriate weight for the fraud data attributes. Therefore, our experiment has shown that for higher fraud detection to be achieved in this kind of data, the weight needs to be set at 0.9.

REFERENCES

[1] Hindu group of publications. (2008, January 23,). $176-b loss due to telecom fraud. Business Daily. Available: http://www.thehindubusinessline.com/2006/09/08/stories/2006090803460400.htm

[2] K. G. Rupesh, and Saroj, K. Meher, "Rule-based Approach for Anomaly Detection in Subscriber Usage Pattern " World Academy Of Science, Engineering And Technology, 2007.

[3] C. Phua, et al., "A comprehensive survey of Data Mining-based Fraud Detection Research," Artificial Intelligence Review, 2005.

[4] P. Burge, and S. John, "An unsupervised neural network approach to profiling the behavior of mobile phone users for use in fraud detection, Journal of Parallel & Distributed Computing vol. 61, pp.915-925 2001.

[5] G. L. Kamp, Steffen and Globig, Christoph, Related Areas. London, UK: Springer-Verlag, 1998.

[6] R. K. Gopal, S. K. Meher,, "A Rule-based Approach for Anomaly Detection in Subscriber Usage Pattern," in Proceedings of World Academy of Science, Engineering and Technology, 2007, pp. 396-399.

[7] J. McGibney, and S. Hearne, , "An Approach to Rules Based Fraud Management in Emerging Converged Networks," presented at the IEI/IEE Irish Telecommunications Systems Research Symposium (ITSRS), Republic of Ireland, 2003.

[8] T. Fowcett, and F., Provost, "Combining Data Mining and Machine Learning for Effective Fraud Detection," AI approach to fraud detection and Risk Management, Workshop Technical Report WS-97-07, pp. 14-19, 1997.

[9] M. Taniguchi, M., Haft, J., Hollmen, and V., Tresp, "Fraud detection in communication networks using neural and probabilistic methods, " in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on Seattle, WA , USA 1998, pp. 1241-1244.

[10] R. J. Frank, N.Davey, and S.Hunt, , "Applications of neural networks to telecommunications systems," in European Congress on Intelligent Techniques and Soft Computing (EUFIT'99), 1999, pp. 255-259.

[11] A. Aamodt, and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," AI Communications, vol. 7, pp. 39-59, 1994.

[12] R. Wheeler and S. Aitken, "Multiple algorithms for fraud detection," Knowledge-Based Systems, vol. 13, pp. 93-99, 2000.