A Fraud Detection Approach in Telecommunication using Cluster GA

V.Umayaparvathi
Dept of Computer Science, DDE, MKU

Dr.K.Iyakutti
CSIR – Emeritus Scientist, School of Physics, MKU

Abstract:

In trend mobile is one of the important devices in public sector. In mobile community there are N numbers of vendors, manufacturers available. The telecommunications industry was first to adopt data mining technology. Telecommunication companies day by day generate and store enormous amounts of high-quality data, have a very large customer data base and operate in a rapidly changing and highly competitive business environment. Telecommunication companies utilize data mining technique to improve their marketing efforts, identify fraud and better manage their telecommunication network connection. In this paper we provide an effective solution to identify the fraud detection in telecommunication using Data Mining clustering techniques with GA as well as misbehavior users. We believe this approach will help in our society.

Keywords: Sequential pattern clustering, Genetic Algorithm (GA), Data Mining, Telecommunication.

Introduction:

The international telecommunications play an important role in message sharing and information passing. In the last decade a dramatic change in the structure of telecommunications companies has been taken place, from public monopolies to private companies. The quick development of mobile telephone networks and video calling and Internet technologies has created enormous competitive pressure on the companies sector. As new competitors arise in market, telecom need intelligent tools to gain profit and withstand. Also, stock market expectations are huge and investors, financial analysts need tested tools to gain information about how companies perform financially compared to their competitors, what they are good at, who are the major competitors are, etc. In other term, the telecom companies need to benchmark their performances against compete trends in order to remain important role in this market. There is a enormous amount of information about these companies financial performance that is now publicly available. This amount greatly exceeds our capacity to analyze it; the problem is that we often lack tools to quickly and accurately process these data.

Data mining for telecommunications companies involve the use of simple, traditional and advanced mathematical techniques used to analyze large populations of data and deliver insights, forecasts, explanations and predictions of how systems, customers, network and marketplace are likely to react to different situations. The hypercompetitive nature of the telecom industry has created a need to understand customers, to keep them, and to model effective ways to market new products. This creates a great demand for innovation like data mining technique to help understand the new business trends involved, catch fraudulent activities, identify telecommunication patterns, make better use of resources and improve the quality of services.

Types of telecom data

The Initial step in the data mining process is to understand the data. Here we discuss three main types of telecom data.

- **call summary data**
  Every time a call is placed on a telecom network, descriptive information about the call is saved as a call detail for future record. At a minimum, each call detail record will include the originating and terminating phone numbers, the date and time of the call and the duration of the call.
- **Network data**
  Telecommunication networks are extremely complex configurations of equipment, comprised of interconnected components. Each network element is capable of generating error and status messages, which leads to a tremendous amount of network data.
- **customer data**
  Telecommunication companies, like other businesses have millions of customer's. For necessity they have to maintaining a database of information on these customers. this information will include name, address and may include other information such as service plan, credit score, contract information, family income and payment history.

Data mining applications in telecom

The telecommunication industry was an early adopter of data mining technology and therefore many applications exist. Four major applications include: marketing customer profiling, fraud detection, churn management and network fault isolation. Each one will be described in more detail. Also many other applications exists which can be classified in these four categories such as : telecom market research , telecom customer segmentation , telecom market sizing , telecom territory design and alignment , telecom market forecasting , telecom fraud prediction , telecom sales force optimization and telecom marketing campaign optimization.

Data mining costs

Unfortunately, telecom data mining tools and algorithm makings are very expensive proposition. The big costs are associated with finding the data cleansing, required it and making it useful. Privacy concerns are an important issue for data mining especially in telecom industry, since telecom companies maintain highly private information, such as which each customer calls. Also there are many legal restrictions in each country about data mining, much of the rationale for these prohibitions relates to competition. The aim of this paper is to find out the fraud detection in telecommunication using Data Mining clustering techniques with GA.

Related Works:

There are several algorithms for mining sequential patterns. Apriori, GSP, SPADE, Prefix Span and Spam are just the simpler ones. From these, GSP and Prefix Span are the best known algorithms, and represent the two main approaches to the problem: apriori-based and pattern growth methods. Next, we will describe both approaches and compare their advantages and disadvantages. Apriori based Approaches. *GSP* follows the candidate generation and test philosophy. It begins with the discovery of frequent *1* sequences, and then generates the set of potentially frequent (*k+1*) sequences from the set off frequent *k*-sequences. The generation of potentially frequent *k*-sequences uses the frequent (*k-1*) sequences discovered in the previous step, which may reduce significantly the number of sequences to consider at each moment. Note that to decide if one sequence *s* is frequent or not, it is necessary to scan the entire database, verifying if *s* is contained in each sequence in the database. In a way to reduce its processing time, *GSP* adopts three optimizations. First, it maintains all candidates in a hash-tree to scan the database once per iteration. Second, it only creates a new *k* candidate when there are two frequent (*k-1*) sequences with the prefix of one equal to the suffix of the other. Third, it eliminates all candidates that have some non frequent maximal subsequence. By using these strategies, *GSP* reduces the time spent in scanning the database, increasing its general performance. In general, apriori based methods can be seen as breath-first traversal algorithms, since they construct all k patterns simultaneously.

At each step *GSP* only maintains in memory the already discovered patterns and the *k* candidates. Pattern growth methods are a more trendy approach to deal with sequential pattern mining problems. The key idea is to avoid the candidate generation step altogether, and to focus the search on a restricted portion of the initial database. Prefix Span is the most promising of the pattern-growth methods and is based on recursively constructing the patterns, and simultaneously, restricting the search to projected databases. A projected database is the set of subsequences in the database, which are suffixes of the sequences that have prefix a. At each step, the algorithm looks for the frequent sequences with prefix a, in the corresponding projected database. In this way, the search space is reduced at each step, allowing for better performances in the presence of small support thresholds.

In general, pattern growth methods can be seen as depth first traversal algorithms, since they construct each pattern separately, in a recursive way. As pointed in, when a gap constraint is used, neither Prefix Span nor Prefix Growth can be applied directly. The generalization proposed Gen Prefix Span, is based on the redefinition of the method used to construct the projected databases. Instead of looking only for the first occurrence of the item, every occurrence is considered.

Proposed Method:

Sequential Pattern Mining algorithms solve the problem of discovering the maximum frequent sequences in a given database. Algorithms for this problem are relevant when the data to be mined has some sequential nature, when each piece of data is an ordered set of elements, like events in the case of temporal information. The goal of sequential pattern mining is to discover all frequent sequences of item sets from the dataset. In particular, an item set is a non empty subset of elements from a set the item collection, called items. In this manner, an item set represents the set of items that occur together. The item set composed of items a and b is denoted by ab. A sequence is an ordered list of item sets. A sequence is maximal if it is not contained in any other sequence. A sequence with *k* items is called a *k* sequence. The number of elements in a sequence *s* is the length of the sequence and is denoted by |*s*|. The *i*th item set in the sequence is represented by *s*i and the set of considered sequences is usually designated by database (DB), and the number of sequences by database size.

Sequential Pattern Clustering Definition:

Let I ={x1...xn} be a set of items. An item set is a non-empty subset of items, and an item set with k items is called k-item set. A sequence s=(X1...X*l*) is an order list of item sets, and an item set Xi ($1\leq i \leq l$) in a sequence is called a transaction. In a set of sequences, a sequence s is maximal if s is not contained in other sequences.

Genetic Algorithm:

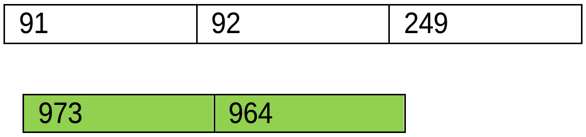
Genetic algorithms are one of the best way to solve a problem for which little is known. They are a very general algorithm and will work well in any search space. All you need to know is what you need the solution to be able to do well, and a genetic algorithm will be able to create a high quality solution. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem.

Genetic algorithms tend to thrive in an environment in which there is a very large set of candidate solutions and in which the search space is uneven and has many hills and valleys. True, genetic algorithms will do well in any environment but they will be greatly outclassed by more situation specific algorithms in the simpler search spaces. Therefore you must keep in mind that genetic algorithms are not always the best choice. Sometimes they can take quite a while to run and are therefore not always feasible for real time use. They are, however, one of the most powerful methods with which to quickly create high quality solutions to a problem. Now, before we start, I'm going to provide you with some key terms so that this article makes sense.

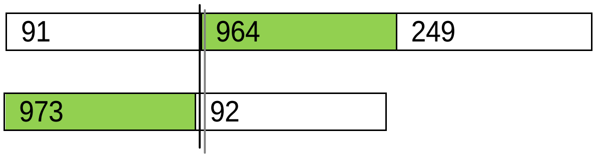
Sequential pattern clustering with GA:

We are going mingle with sequential pattern clustering with Genetic algorithm to optimize the SPC (sequential pattern clustering). In this work telephone no code act as chromosome Based on that we can generate the initial population size from that we can identify the group of corresponding number groups. In the large no group we will apply the GA for generating the new optimized group based on some fitness function value. Fitness function value is nothing but it's a criterion for generating the new solution or new candidate functions. Before get into the technique we have to follow some initial pre activities like defining the fitness value, initial population size, mutation point, crossover point and ordering data. After define everything we will apply the technique into the data.

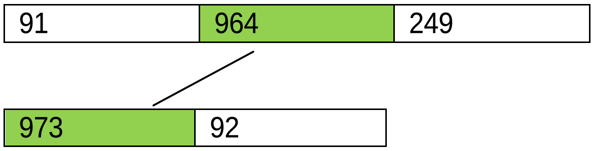
In our telecommunication data base contain 2000 records, which are got from one of the Indian call center sector. Next thing what we going to define is initial population size that's near to 1000 records in that we are going to apply the genetic algorithm(GA) technique. In GA there are two important concepts is there like mutation, crossover. We will see how this technique will work.



Samples



Crossover



Mutation

Input:

```
Initial samples size: N
Maximum generations: G
Threshold value: T
Minimum fitness: minF
Output:
Malpractice user no lists
Begin
Step1: Initialize counter count = 0
Step2: Initial population IN of size N
Step3: For each chromosome i ∈ IN
If S1& S2 is given then
Measure the fitness F(i, S1, S2,T)
Else If S1 is given then
Measure the fitness F(i, S1, T)
Else If S2 is given then
Measure the fitness F(i, S2, T)
Else
Measure the fitness F(i, T)
Step4: Mutate and crossover P.
Step5: IF (fitness ≥ minF)
Select fittest rules from P
Step6: Set temp = temp +1
Step7: IF (t > G) then
S = P
Stop
Else
Go to Step 3
End
```

Fitness function measurement will evaluate the sequential pattern clustering. Based on that we will fix threshold value to determine the spoofed calls and malpractice customer no. threshold value determinate the whether the call is legitimate or not. Its value derived based on the time duration of the call, country code, destination and frequency. Destination code acts as important point in threshold value because customer can modify or clone them mobile no but destination no won't change at any cost. It's tedious process to identify which customer legible or hacker. Even though they change the IMEI no we can identify the based on the destination no and country code.

Result Analysis:

S.No	Phone number	Date & Time	Destination	Country code	Call duration
1	**446215882**	14-12-2008 18:46:52	63	92	00:01:48
2	**446215917**	14-12-2008 21:25:45	63	92	00:00:41
3	**446215966**	14-12-2008 17:30:51	63	92	00:02:47
4	**446212232**	13-12-2008 05:55:31	973	92	00:01:21
5	**446212296**	14-12-2008 05:32:59	973	92	00:01:25

This simulation done in mat lab editor while executing this algorithm we will get list of customer those who are consider as malpractice user or hacker. The sample out put displayed here instead of complete model. Some time our algorithm results legitimate customer as hacker or malpractice person. We have to choose exact person no who is hacker because its just list out all possibilities of hacker no.

Conclusion:

In this paper we provide an effective solution to identify the misuse customer and malpractice customer in the telecommunication area using sequential pattern clustering and genetic algorithm. This paper will help to society for utilizing mobile in secure way and identify the misbehavior customer in telecommunication department.

Future work:

In future we are going to develop model for business trends, marketing strategy, analysis report from the existing database. Some time our algorithm result legitimate customer as hacker or malpractice person. We have to overcome this in our future work. In this paper we provide an effective solution for identify the fraud and malpractice user identification. In future we going to develop the complete model for telecom industry to enhance and compete with other vendors.

References:

[1] Saleh Al Kodhair, Mining Association Rules using Genetic Algorithm, 2008, Master Thesis.

[2] S. Sakurai, Y. Kitahara, and R. Orihara, "A Sequential Pattern Mining Method based on Sequential Interestingness", Fall.2008, International Journal of Computational Intelligence, pp.252-260.

[3] Jay Ay res , Johannes Gehr ke, Tomi Yiu, and Jason Flannick,"Sequential Pattern Mining using A Bitmap Representation", 2002, SIGKDD '02 Edmonton, Alberta, Canada.

[4]Malone, J., etc.. (2005). Data mining Using Rule Extraction from Kohonen Self-organising Maps. Neural Comput & Applic 15: 9-17.

[5]Wu, S., Gao, X. and Bastian, M. (2003). Data Warehousing and Data Mining. Metallurgical Industry Press. Beijing.

[6]Wang, Y. (2004). The Study on Data Warehouse and Data Mining in Telecom industry Management Analysis and Application. College of Computer Science, Chongqing University. Chongqing.