An Approach for Effective Use of Pattern Discovery for Detection of Fraudulent Patterns In Railway Reservation Dataset

Rasika Ingle[1], Manali Kshirsagar[2]
[1] Dept. of Computer Technology,Yeshwantrao chavan college of Engineering,
Nagpur,Maharashtra, India
[2]Dept. of Computer Technology,Yeshwantrao chavan college of Engineering, Nagpur,Maharashtra,India

Abstract:

Data mining concepts and techniques can help in solving many problems. Useful knowledge may be hidden in the data stored. This knowledge, if extracted, may provide good support for planners, decision makers, and legal institutions or organizations. Hence Pattern discovery, as one of the powerful intelligent decision support platforms, is being increasingly applied to large scale complicated systems and domains. It has been shown that it has the capacity to extract useful knowledge from a large data space and present to the decision makers. This will contribute to the detection of illegal activities, the governance of systems, and improvements in systems. This paper proposes a work to develop a mechanism that allows the system to work interactively with a user in detecting, characterizing and learning unusual and previously unknown patterns over groups of records depending on the characteristics of the decisions. The data mining in real time could even help to alert Railways when something untoward happens. Hence this innovative mechanism focuses on detecting anomalous and potentially fraudulent behavioral patterns within set of railway reservation transactional data .The pattern based analysis will include the possible detection of fake ids, fake booking , an unusual pattern like reservation of a person for trains in two different directions on a given date form the same starting city etc.

Keywords: Anomalous transactions, data mining, fraudulent transactions, hash map, pattern, pattern discovery, rule based discovery.

1. Introduction

All Data Mining is the process of discovering new correlations, patterns, and trends by digging into (mining) large amounts of data stored in warehouses, using artificial intelligence, statistical and mathematical techniques. Data mining is the principle of sorting through large amounts of data and picking out relevant information. It has been described as "finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning" [13] and "the science of extracting useful information from large data sets or databases". The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. According to this view, data mining is only one step in the entire process, albeit an essential one because it uncovers hidden patterns for evaluation [14]. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. One of main area where data mining can be used in the industry is in monitoring systems. The specific tasks in automated transaction monitoring systems are the identification of suspicious and unusual electronic transactions. An unusual pattern is an observation or a point that is considerably dissimilar to or inconsistent with the remainder of the data. Detection of such outliers or patterns is important for many applications and has recently attracted much attention in the data mining research community.

Pattern-based analysis looks for anomalies indicative of fraud or error in normal patterns of data. It is growing gradually and becomes more important with the quick development of computer technologies with increasing capacity to collect massive amounts of valuable data for pattern analysis. In real life, fraudulent transactions are interspersed with genuine transactions and simple pattern matching is not often sufficient to detect them accurately. Often times, discrepancies in transaction data are missed when analysis doesn’t go beyond known problems. Discrepancies may result from unanticipated behavior that pattern-based analysis is more apt to uncover.The basic question asked by all detection systems is whether anything strange has occurred in recent events. This question requires defining what it means to be recent and what it means to be strange.“What’s strange about recent events”. WSARE operates on discrete data sets with the aim of finding rules that characterize significant patterns of anomalies [9]. In general, anomalies can be defined as any observations that are different from the normal behaviour of the data. Many traditional anomaly detection techniques look at the data records individually, and try to determine whether each record is anomalous with respect to the historical distribution of data. A Bayesian Network likelihood model and a conditional anomaly detection method are considered by [10],[11]. In terms of data mining, fraud detection can be understood as the classification of the data. Research on fraud detection has been focused on the pattern matching in which abnormal patterns are identified from the normality [12]. Input data is analysed with the appropriate model and determined whether it implies any fraudulent activities or not. One could use the classical data mining tools to get supporting data to confirm or refute existing personal perception, but one also cannot be assured that there are no better-fitting explanations for the discovered patterns, or even that no important information has been missed in the entire data mining process. For a relatively complex real problem with a large data space, all traditional knowledge acquisition and data mining tools would become obviously inefficient, even helpless in some ways. For a larger mixed-mode database with more unanticipated variations than normal ones, even the domain experts would find it difficult to reach useful results. Hence this limitation motivates to develop a technique that searches for suspicious patterns in the form of more complex combinations of transactions and other evidence using background knowledge. There are many indicators of possible suspicious (abnormal) transactions in traditional illegal business. Here we concentrate on fraudulent patterns; This paper proposes a work that aims to identify certain forms of knowledge that can be inferred from the information infrastructure that supports railway reservation booking systems. It will also assess the integration of data mining with these systems as a means to facilitate the extraction of useful knowledge.The purpose of this short paper is to present an idea of finding fraudulent patterns from a railway reservation dataset and proposing a mechanism to achieve the same. The structure of the rest of the paper consists of an introduction to data mining and pattern discovery concepts, followed by a brief description of the key aspects of proposed work. This section sets the scene for the main methodology for a fraudulent pattern discovery. The paper then concludes with a scope and limitation.

2. Methodology

2.1. Problem Definition

Pattern discovery from large datasets has been an active field of research for the past two decades. These studies are driven by a desire for automated systems which can search, analyse, and extract knowledge from the massive amount of data collected in many fields. The main goal is to replace the conventional manual examination methods which are expensive, inaccurate, error prone and limited in scope. Reservation records should also be searched for unusual patterns and undiscovered knowledge. This proposed work demonstrates that different kinds of illegal manipulation or ways used in railway reservation transactions can be discovered by identifying particular patterns and track them in the datasets. Fraud indicators in the railway reservation transactions are the focus of this work. The problem is formulated by,

Recognising those indicators, the patterns associated with them, and the human behaviour underlying these patterns;
A data mining approach to automate the discovery of the illegal activities that generate the patterns.

2.2. Significance

This study will contribute to current efforts in establishing better systems to support the railway reservation governance. To achieve this, two main problems are addressed.

Assessing the patterns hidden in reservation transaction records which can be used to point out useful knowledge.

Automating the discovery of some of these patterns from reservation records by applying data mining techniques.

A major problem is the lack of published work that addresses the automatic extraction of knowledge from railway reservation systems. Therefore, in some aspects, this is a pioneering study.

2.3. Research Objectives

The primary objective of this work is,
To explore the use of data mining in railway reservation systems and to develop knowledge of where and how data mining can be applied and integrated into these systems, to contribute to the discovery and alleviation of fraud in railway reservation transactions.

As stated, the primary objective of this work is set to provide a solution to fraud bookings by agents and to railway reservation governance by detecting fraud. To serve this primary objective, four main activities or sub-objectives are set. They are,

2.3.1 Identify different fraudulent activities in reservation record datasets, in a variety of contexts where these activities may take place.

2.3.2 Identify suitable data mining techniques that may help in detecting some of the fraud activities found in (2.3.1).

2.3.3 Design and develop a data simulator to generate reservation record datasets.

2.3.4 Identify existing tools or techniques to apply the methods found in (2.3.2) above and develop a mechanism which serves the main objective.

2.4. Suggested Approach

This section of the paper describes the activities that would be carried out by the authors in order to address the problem of the research.

2.4.1 Acquisition of Domain Knowledge

In addition to the literature review; a qualitative survey was conducted using unstructured interviews because literature is sparse. This survey asked experts in the field of railway reservation systems (Railway employees, travel agents and consultants), about their knowledge in indicating the types of frauds or unusual patterns that may occur in reservation records.

2.4.2 Data collection

For the proposed work the relevant data set is the records of railway reservation transactional data. Hence the required data set is collected and studied successfully. It has been studied that this transactional database mainly consist of master files as

- Train Index
- Station Index
- Category Index
- Fare Index
- PNR Index
- Current Index
- Reservation Index

These master files contain the overall details of specific nature. Based on the observations made from the collected railway reservation transactional data set, the required relevant database is created following the same structures.

2.4.3 Creation of fraud schemes based on the findings from the first two activities above

In this activity the author studied and analysed the behaviours found in the first two activities to come up with a set of racketeering methods and schemes that are used in the real world. For each scheme, the author tried to find the effect of conducting it on the datasets and how the corresponding records differ from the records of any other usual reservation activity. The findings are summarized in lists of fraud patterns and indicators. Based on the patterns and indicators studied for the schemes, some of the schemes would be selected for the testing of data mining. This selection process is mostly based on the available data.

2.4.4 Data mining methods used in pattern discovery

This phase encompasses two activities. First performs the study of data mining concepts and techniques [5]. This will help in understanding the problems data mining techniques might be useful in solving. The second activity is to understand the existing fraud detection techniques and examine some case studies of fraud detection in different fields [12]. This is an important activity since the focus of this study is more toward detecting fraudulent activities in railway reservation transactions. This review helps to understand how to formulate schemes found in order to apply detection methods.

2.4.5 Data simulation

In this step, a reservation records simulation system was developed to provide the required data for the experiments in this research. The need for a simulator stems from not only the lack of uniform data sources but also the lack of access to railway reservation bookings data.

2.4.6 Formulation of domain specific algorithms

For each fraud method selected in activity mentioned above (2.4.3) the rule of detecting fraudulent activities in that scheme is formulated. A search is conducted for available tools such as data mining toolboxes provided in MATLAB or the machine learning/data mining software (WEKA) .These tools can provide detection methods that can be applied to extract the targeted pattern from the records. If no appropriate available tools were found, then database queries will be implemented by the author. By analyzing the filtered transactions, a domain specific rule based algorithm will be designed for finding abnormal transactions at this step.

2.4.7 Testing

Perform the tests for the developed mechanism on the appropriate datasets and analyse the results. This step includes performance evaluation for the used methods.

3. Scope And Limitation

Data mining concepts and techniques can help in solving many problems. This paper suggests that data mining should be studied and applied in railway management; however, it only investigates the application of a set of techniques that are used for fraud detection. Furthermore, it is not the purpose to compare and evaluate performance and efficiency of different algorithms; the main goal is to evaluate the usability and the efficiency of data mining algorithm in the context of the pattern discovery. The reason of this scoping of the current work is because it is pioneering work and more focused on addressing the problem and the solutions while efficiency can be developed later. One of the major limitations of this study is the lack of real datasets. As data is a major factor in the success of achieving the objectives, the existence of real datasets would have helped substantially in the progress of the research. A property transactions data simulator was developed to overcome the lack of data availability. While this simulator has advantages, it has some limitations and creates a finer scope for the type of data to be worked with.

References

[1] Junjie Wu, Shiwei Zhu, Hui Xiong,Jian Chen, and Jianming Zhu, “Adapting the Right Measures for Pattern Discovery: A Unified View”, IEEE Trans. On Systems ,Man and Cybernetics-Part B ,Vol.42,No.4,Aug 2012.

[2] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,” Effective Pattern Discovery for Text Mining”, IEEE Trans. Knowledge Data Eng.,Vol. 24, No. 1, Jan 2012.

[3] Mehmet Koyuturk, Ananth Grama, and Naren Ramakrishnan , “Compression, Clustering, and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets”, IEEE Trans. Knowledge Data Eng.,VOL. 17, NO. 4, APRIL 2005.

[4] Dongsong Zhang and Lina Zhou ,“Discovering Golden Nuggets: Data Mining in Financial Application” IEEE Trans. On Systems ,Man and Cybernetics-Part C:Application and Reviews , vol. 34,no. 4,Nov 2004.

[5] Kovalerchuk, B., Vityaev, E., “Detecting patterns of fraudulent behavior in forensic accounting”, In Proc. of the Seventh International Conference “Knowledge-based Intelligent Information and Engineering on Systems”, Oxford, UK, part 1, pp. 502-509, Sept, 2003.

[6] Andrew K. C. Wong, Senior Member, IEEE, and Yang Wang, Member, IEEE, “Pattern Discovery: A data driven approach to decision support”, IEEE Trans. On Systems , Man and Cybernetics-Part C:Application and Reviews, vol. 33,no. 1,Feb 2003.

[7] T. Chau and A. K. C.Wong, “Pattern discovery by residual analysis and recursive partitioning,” IEEE Trans. Knowledge Data Eng., vol. 11, pp.833–852, Nov./Dec. 1999.

[8] Nitin Jindal, Bing Liu, Ee-Peng Lim, “Finding Unusual Review Patterns Using Unexpected Rules”.

[9] Weng-Keen Wong,Andrew Moore,Gregory Cooper, and Michael Wagner,”Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks”.

[10] Kaustav Das,Jeff Schneider, Daniel B.Neill, “Anomaly Pattern Detection in Categorical Datasets”.

[11] Kaustav Das, Jeff Schneider,”Detecting Anomalous Records in Categorical Datasets”.

[12] Jia Wu and Jongwoo Park ,” Intelligent Agents and Fraud Detection”.

[13] Margaret H. Dunham, Data mining: Introductory and advanced topics,Dorling Kindersley (India) Pvt. Ltd.,Pearson,2006.

[14] Jiawei Han , Micheline Kamber , Data mining: Concepts and Techniques, M Morgan Kaufmann , 2005.