Pattern Behaviour Prediction Using Data Mining in Insurance Application

V. Saillaja[1*], Senthil Kumar Seeni[2]

[1] *Department of Management Studies, Sri Sairam Engineering College, Chennai, Tamil Nadu, India.*
[2] *Department of Mobile Application Development, Cognizant Technology Solutions,*
Buffalo Grove, Illinois, United States of America.

**Corresponding author: psaillaja@gmail.com*

Abstract. Gigabytes of data gathered by the health insurance sector have been analysed and mined to see if two data mining approaches are effective in uncovering previously undisclosed behavioural patterns. The act of choosing, examining, and modelling vast volumes of data to reveal previously unknown patterns is known as data mining. In the insurance sector, data mining may assist companies in gaining a competitive edge. The act of choosing, examining, and modelling vast volumes of data to reveal previously unknown patterns is known as data mining. In the insurance sector, data mining may assist companies in gaining a competitive edge. Both incident databases for diagnostic services as well as a general practitioner's database were used in this study. The event database was analysed using association rules, and neural segmentation was used while overlaying the two datasets. It was shown that data mining in healthcare insurance information management may be used to identify trends in ordering pathology services and to categorise primary care doctors into groups based on the practise style and type. Conventional methods could not have produced the outcomes that were reached utilising the strategy employed.

Keywords: Data Mining, Neural Segmentation, Association Rules, Dataset Overlay.

INTRODUCTION

Numerous businesses have implemented extensive information systems, which keep detailed records of all kinds of operational activities to better utilise data for planning and strategic company development. Gigabyte databases are getting increasingly popular, and the old query and report-based techniques of data analysis are becoming increasingly ineffective. With data mining, correlations, rules, and functionalities may be presented to a trained person for inspection and scrutiny via automated presentation [1]. Each pattern, rule, and function have an associated domain expert who weighs in on whether these are important to the paradigm. A general pattern recognition model is shown in Figure

Research and choice of stocks, fraud detection, spending patterns, and patterns of bit failures in semiconductor memory manufacture are only some of the uses of data mining techniques that have been widely used in recent years. Due to rising health-care expenditures and the increasing need to keep these expenses under control, timely evaluation of health-care data has become a critical concern. Health-care data analysis is a time-consuming, expensive endeavour for large firms, hospitals, health-care management groups, and insurance companies. Many companies have started to build automated systems to help with the process. Travellers Insurance, for example, has created a method for identifying healthcare provider fraud in electronically submitted claims [2].

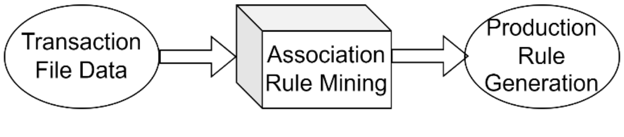

FIGURE 1. General Pattern Recognition Model

Received: 03.02.2021 Revised: 04.03.2021 Accepted: 22.03.2021

An online claim file alone has more than 500 GB of data, dating back to the last five years. There has been a lot of work done to ensure that the data used in developing and implementing methods to identify and avoid corruption and unethical behaviour is as accurate and relevant as possible. To analyse claims data, it is necessary to identify and construct acceptable utilisation patterns and linkages. A lot of data analysis has focused on fraud detection and prevention to this point. Unreasonable, needless, or excessive service requests or services are examples of inappropriate practise. For this sort of study, human expertise is necessary yet expensive and in short supply. However, this technique can only be utilised on a portion of the information, which is why HIC has incorporated a neural network to help with the process [3].

EXISTING WORK

Segmentation, extrapolation, grouping, synthesis, and dependency modelling are some of the most used data mining techniques. An association rule summary approach and a neural segmentation implementation of clustering are discussed in this study. Following is a summary of the strategies the problem can be solved, as well as a brief discussion of the fundamental characteristics. The challenge of applying association rules to basket data transactions has been identified. The definition of association rule mining, which is used to look for patterns within transactions [4]. In the case of a database containing transactions, where each transaction represents a collection of objects (such as services performed), find all possible connections between the various items in the database. Data files, relational tables, or the output of a relational expression can all be used as databases. As the size of the database grows, the time it takes to process it increases linearly with the volume of data. During this research, these traits have held true. Using a self-organizing feature map as the foundation, neural segmentation may find patterns in images.

A self-organizing feature map, also known as a topological feature map, is a two-dimensional representation of a multidimensional space. In other words, comparable prototypes are clustered together on the map. Units in a self-organizing feature space are coupled to n input nodes, and each unit is identified by an n-dimensional vector that specifies its placement in the array [5]. Euclidean distance between input and weight is computed by each neuron. The input pattern is assigned to the neuron that has the shortest distance between it and the other neuron. The depiction of behavioural data is critical to the effectiveness of this study. Care must be made to ensure that the inputs are evenly distributed in the self-organizing feature maps since clustering is a kind of self-organization [6]. To put it another way, each vector element reflects one equally balanced dimension of the subject's behaviour inherently." To maintain an even number of vector components across all the inputs, it is necessary to balance the inputs. For this, a thorough investigation of the problem is required. This includes looking closely at the kind of features that are critical to the solution and closely examining the variance in each of the variables [7].

PROPOSED SYSTEM

The database was pre-processed such that only the characteristics to which connections would be applied could be extracted from the database. This procedure is dependent on a tool that makes it possible to efficiently manipulate millions of data [8]. Diagnostic treatments or medical services are of interest in this study. A transaction id and n characteristics are needed in the record format specified by the association rules algorithm. A transaction-id/attribute pair was used to arrange the data in the episode database schema specification since there are only 20 unique characteristics per event [9]. The patient-id as well as time stamp were used to generate a unique transaction-id, which was then appended to each record when the characteristics were extracted from the database [10]. Using the Apriori method, association rules it can be derived Names files with the text descriptions of each code utilised in the transaction file are also included in the inputs. The episode database for pathology services was subjected to neural segmentation. Doctors ordering profiles were examined to identify subgroups within the profile population based on the kind of tests it is requested.

The doctor database was used to gather information on the practice's characteristics, as well as the kind of tests ordered and the frequency with which it was administered [11]. The proposed system is shown in Figure 2. The episode database was utilised to generate the segmentation algorithm's input vectors. Because each episode might include up to 20 tests, the database entries must be rotated and aggregated by GP. Scaled depiction of the probability of every test requested by the GP throughout a whole year was constructed using 10,409 vectors [12]. According to the total series of experiments, the number of tests done by each doctor was scaled from 0 to 1. The vectors were presented to the system for 25 epochs (also known as iterations) to train it repeatedly. An initial learning rate of 0.8 was utilised, with the end value of 0.05; the learning rate was lowered linearly after each epoch.

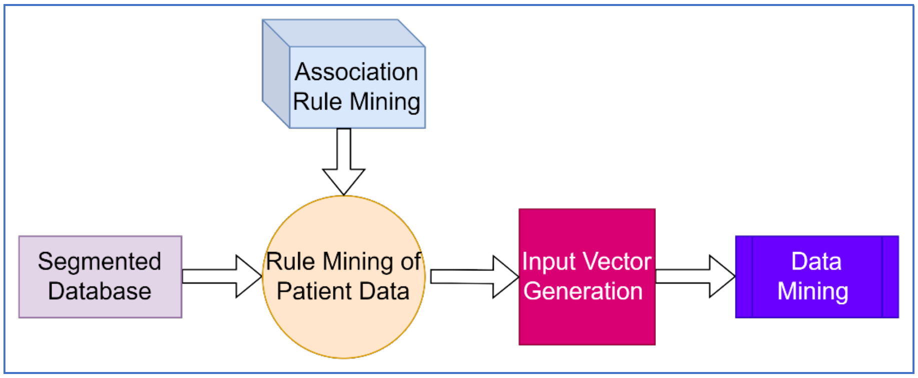

FIGURE 2. Proposed System Architecture

The breadth of the Gaussian distribution function (GDF) was changed from the ratio of the number of endpoints to 0.1 in a neighbourhood function. The episode database for pathology services was subjected to neural segmentation [13]. To identify subpopulations within that profile based on the kind of tests that are requested this must be looked at how general practitioners order the tests. The practitioner database was utilised to gather information on the practice's nature and to identify the selection and frequency of tests [14].

RESULTS AND DISCUSSION

The user was able to specify 1 percent, 0.5 percent, and 0.25 percent (minimum support) to produce association rules with a confidence level of 50%. There have been 68,000 transactions in pathology services out of a total of 6.8 million episodes. Each experiment yielded a different number of association rules [15]. The algorithm came up with simple and sophisticated criteria, and the most common medical test is performed in 28.15 percent of all events. A collection charge accounted for an extra 20.15 percent of the increase. 48.30 percent of pathology service sessions had this test present [16].

A total of 10.9 percent of all episodes included the most often requested combination of diagnostic tests. There was a 62.2 percent likelihood that if test-A was requested and then test-B would be ordered as well, according to the rules provided. The examples above demonstrate a significant number of episodes in which common tests are placed together. Consideration must be given to the propriety of arranging combinations such as those shown below. During these investigations, it appears that a substantial number of pathology tests are ordered based on a screening method rather than a pre-planned strategy [17]. It's possible that association rules and more extensive data mining can provide a way for medical learned bodies to deliver more broad and targeted instruction. The numerical results of execution time analysis are listed in Table 1.

TABLE 1. Execution Time Analysis

Threshold	CPU Time	Elapsed Time
1	25.68	32.23
0.5	25.79	33.36
0.25	26.21	33.65

The significance of association rules in revealing otherwise missing associations is demonstrated by new and unexpected findings found on test-X. Higher-level views of providers' ordering behaviour were acquired through traditional query techniques, which were previously unavailable. Execution time analysis is depicted in Figure 3.

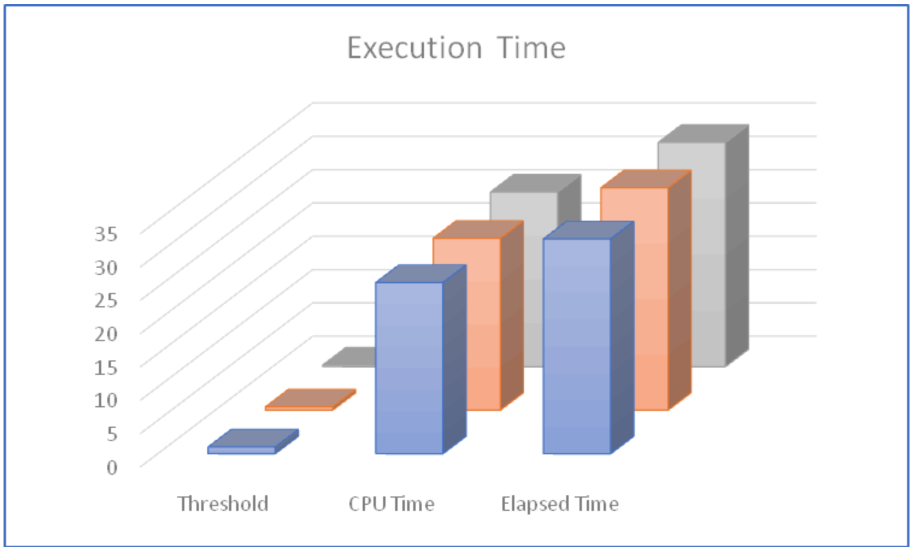

FIGURE 3. Execution Time Analysis

Non-meaningful rules as well as noise in the results from various permutations of collection-fee items and medical services hinder key conclusions and complicate analysis of the output. Considering that 30.20 percent of all events include a collection-fee element, research was carried in which all such items were excluded from the pathologic episode data source. Test results obtained from segment dataset is shown in Table 2.

TABLE 2. Test with Segment Dataset.

Threshold	1	0.5	0.25
Rules Generated	35	68	92

Filtering the collection-fee elements improved the association's rules significantly. According to the newly established standards, both service ordering behaviour and billing practises may now be seen at a high level. After selecting the collection-fee items, an odd combination linked to test-X was readily obvious from the association criteria.

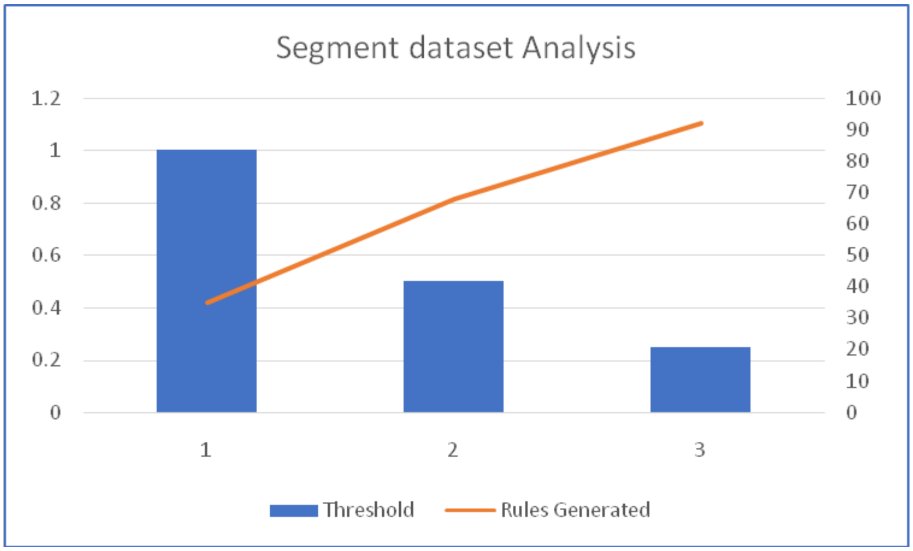

FIGURE 4. Segment Dataset Analysis.

Due to the excellent clinical potential of doing these two tests combined, there are no restrictions for claiming both items. But the data show that test-X is being mistakenly claimed rather than test-Y in a significant number of instances. Whether similar claims practises occur in other jurisdictions is an open subject. Over the past five years, this issue has gone unnoticed by conventional monitoring methods. In Figure 4, segment dataset analysis is depicted.

CONCLUSION

In this study, the efficiency of two data mining approaches in the context of a big health insurance database is investigated. Data mining techniques may be effectively applied has been demonstrated that is huge, real-world customer datasets, with an acceptable execution time, in this study. The findings suggest that these algorithms may assist organisations identify actions that need to be performed, and that this can have quantitative advantages for the business. The findings of the investigation have given to the health insurance company fresh and surprising insights into the mix of services supplied by pathologists. Conventional monitoring methods had been unable to detect this issue, which had been going on for over five years. Based on the practises' nature and style, general practitioners were classified of varying sizes in the study. In the past, categorization of general practise into its different divisions had been difficult. The expanded subgroups will make it possible to keep a much closer eye on how different practises are being used. Neural segmentation gives the tools to comprehend and monitor the behaviour of the various subgroups in an application such as general practise, which is the commercial purpose of health insurance organisations to change practitioners' behaviours towards optimal practise.

REFERENCES

[1]. C. F. Chien and L. F. Chen, 2008, “Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry,” *Expert Systems with applications*, **34(1)** pp. 280-90.

[2]. A. S. DeNisi and R. W. Griffin, 2005, “Human resource management,” *Dreamtech Press*.

[3]. J. Han, J. Pei and M. Kamber, 2011, “Data mining: concepts and techniques,” *Elsevier*.

[4]. J. Ranjan, D. P. Goyal and S. I. Ahson, 2008, “Data mining techniques for better decisions in human resource management systems,” *Int. J. of Business Information Systems*, **3(5)** pp. 464-81.

[5]. M. J. Huang, Y. L. Tsou and S. C. Lee, 2006, “Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge,” *Knowledge-Based Systems*, **19(6)** pp. 396-403.

[6]. G. K. Tso and K. K. Yau, 2007, “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks,” *Energy*, **32(9)** pp. 1761-1768.

[7]. K. Y. Tung, C. Huang, S. L. Chen and C. T. Shih, 2005, “Mining the Generation Xers' job attitudes by artificial neural network and decision tree—empirical evidence in Taiwan,” *Expert Systems with Applications*, **29(4)** pp. 783-794.

[8]. W. S. Tai and C. C. Hsu, 2012, “A realistic personnel selection tool based on fuzzy data mining method,” *9th Joint Int. Conf. on Information Sciences (JCIS-06)* pp. 190-193.

[9]. M. S. Patil and S. Chavan, Evaluation of Data Mining applications in Insurance Sector.

[10]. V. S. Rao and M. V. Jonnalagedda, 2012, “Insurance Dynamics–A Data Mining Approach for Customer Retention in Health Care Insurance Industry,” *Cybernetics and Information Technologies*, **12(1)** pp. 49-60.

[11]. P. Patel, S. Mal and Y. Mhaske, 2019, “A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques,” *Int. Res. J. of Eng. and Tech. (IRJET)* **6(01)**. pp. 591-594.

[12]. M. S. Shah, Pattern Analysis and Selection of Mediclaim Policies Using Data Analytics.

[13]. V. Bhatnagar and J. Ranjan, 2011, “Time to implement data mining in insurance firms for effective CRM and CRM analytics,” *Int.l J. of networking and virtual organisations*, **9(1)** pp. 1-24.

[14]. H. L. Yang and C. S. Wang, 2008, “Locating online loan applicants for an insurance company,” *Online Information Review*, **32(2)**, pp. 221-235.

[15]. P. Sehgal, S. Gupta, D. Kumar and H. PPIMT, 2012, “Application of neural networks in predictive data mining for insurance,” *Int. J. of Latest Trends in Eng. and Tech.*, **1(1)**, pp. 1-4.

[16]. L. Goleiji and M. J. Tarokh, 2016, Survey of Detecting Fraud in Automobile Insurance Using Data Mining Techniques. *Int. J. of Computer & Information Technologies (IJOCIT)*, **4(4)**, pp. 1-10.

[17]. S Murugan, S. Mohan Kumar, and T.R. Ganesh Babu, 2020, “CNN model Channel Separation for glaucoma Color Spectral Detection,” *Int. J. of MC Square Sci. Res.* **12(2),** pp. 1-10.