Anticipate Pattern Mining and Temporary Data Features Extraction in Medical Care System - A Study

[1]P.Sathya, [2]Vignesh Ramamoorthy.H, [3]Juliet Rozario
Assistant Professors,
Department of Computer Science, Sree Saraswathi Thyagaraja College

Abstract **- An important goal of knowledge discovery is the search for patterns in the data that can help explaining its underlying structure. To be practically useful, the discovered patterns should be novel and easy to understand by humans. In this work the problem of mining patterns (defining sub populations of data instances) that are important for predicting and explaining a specific outcome variable. We propose and present efficient methods for mining predictive patterns for both a temporal and temporal (time series) data. Our first method relies on frequent pattern mining to explore the search space.**

Keywords: patterns, data instances, discovery, outcome and temporal.

I. BIG DATA - INTRODUCTION

Data size has increased significantly with the advent of today's technology in many sectors such as productivity, industry and science and web application. Some information is structured, semi-structured, and others are structured with other types of data, including documents, posts, images and videos (Hall, 2013). Such a large data researcher is defined as great data (Queen, 2012). The word big data, Garlasu et al. (2013).

As data compressed in large quantities, conventional database tables cannot be configured and made with great direction. Sen and others. (2013) Defines data with measurements beyond the capability of software tools commonly used to capture, manage and process within an acceptable remainder of the time. Based on these benchmarks, you can decide that there are three aspects of the larger data, such as its size and its shape. Until 2003, 5 global data were 5 extrabytes and 2.7 jettabytes (Garlasu et al., 2013). Many companies begin to investigate the use of large data and how to make profits (Qin, 2012).

A. Big Data Platform Requirements

As for data warehouse, web store or any ID platform, infrastructure for large data has distinct needs. To consider all the algorithms of a large data base, it is intended to easily integrate your large data with your company data to allow deep analysis of the compound data set. [5]Data acquisition, data structure and data analysis requirements for a large data infrastructure [5].

B. Get larger data

The acquisition phase is one of the major changes in the infrastructure from the days before the large data. Since large data represents high volume and high-range data streams, the infrastructure needed to support large data acquisition must both reveal data and short and predictable delay in executing short and simple questions; Most transaction modules can handle, often in a distorted environment; And to secure flexible, dynamic data structures. NoSQL databases are typically used to obtain and store larger data. They are dynamic data structures are perfect and very scalable.

C. Organize Big Data

In conventional data warehouse, synchronizing data is called data integration. Being large-scale data, there is a tendency to sorting data in its initial target location, thereby reducing both time and money without moving large data. The infrastructure needed to organize large data must be able to process the process and data on the original storage location; [1]Very high performance (often volume) to deal with large data processing operations; Handle a large variety of data formats from structured configuration.

Hadoop is a new technology that will enable large data volumes to be ordered and executed when managing data collection cluster data. The Hadoop shared file system (HDFS) for example is a long time storage system for web logs. These web logs have changed the browsing behavior (sessions) by making the mapped programs running on the cluster and creating integrated results in the same cluster. These integrated results are then loaded into a function DBMS system.

II. PROBLEM STATEMENT

Worst data quality, such as poor data quality, dirty data, wrong values, wrong values, wrong or invalid values, poor data quality and poor representation in the data model. Synchronize contradictory or redundant data from various sources and formats: multimedia files (audio, video and images), earth data, text, social number, etc .Individuals, organizations and governments are increasing the security and privacy concerns.

Data is unavailable or hard access to data. Effectively extract the amount of data in the performance and scalable databases of data processing protocols. Larger data on logical and mechanical achievements is most part of the complex data width. In this manner, research is focused on improving logical data management and adequacy at the time of data collection, collection, processing, testing, and funding. The large data pump should be reduced to the pile of the data, the large data trial stage system, the test scale access and the value of data assets. It will give a logical, complete; the

temporary and reliable large data experiment and customized data proposal benefit and logistic and innovative achievements help in the selection of revitalization management.

III. Proposed Methodology And Description

Three kinds of user behavior are mined in this phase: application use, smart device use and periodicity of user behavior. When mining application utilization, the application establishment often times utilized applications and application correlation are dissected. The application use is for some time followed. When mining the device use, the mean, fluctuation and auto correlation are computed both for span and interim. This phase gathers adequate information and mines of user behavior on the over three viewpoint.

A. Mining and Predicting Smart Device user behavior

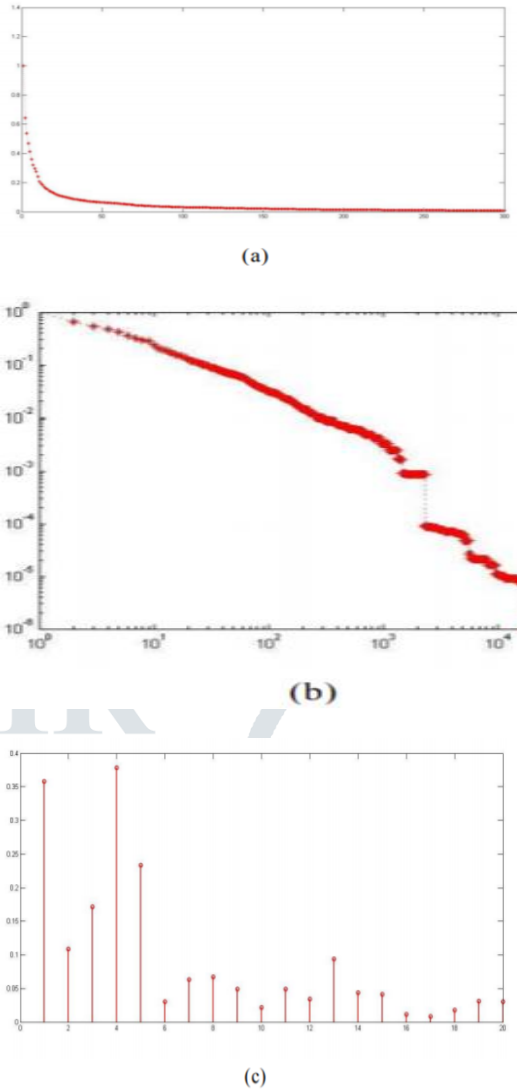
All the customized administrations depend on the understanding of user behavior. Smart device is the 66 closest gears for users, and consequently the mining of smart device usage behavior is the most vital region of mining user behavior, and can contribute much to customized administrations. There are numerous examines on mining portable user behavior, the focal points of these looks into are different. There still need inquires about on the mining of application usage, smart device usage and time highlight of user of behavior. This phase gathers adequate information and mines of user behavior on the over three viewpoint.

B. Device Usage Statistics

First the total time of smart device usage in one day is calculated. In the range of four months, of all the gathered users, the longest time of device usage is 324.2 minutes in a single day, and the briefest usage is 2 minutes, with non-utilize avoided. The normal usage is 53.0 minutes, and standard deviation is 42.4 minutes. So we can see that there are no monster hole between the most dynamic users and the minimum dynamic users. And then the term of device usage at once is examined.

Here span has indistinguishable significance from period in 2.2, amid which the screen is never off. Of the considerable number of lengths, the normal term is 60.9 seconds, the standard deviation is 241.5 seconds, the most extreme is 299.3 minutes, the base is 0.7 seconds and the coefficient of variety is 396.6%. With respect to one user, the term is additionally unique. The CDF (combined conveyance work) of every one of these terms is appeared in Figure 1(a). The capacity in log-log arranges is about direct, as appeared in Figure 1(b). Through the R-square test, the correlation coefficient is 0.9373, so it is inferred that the length of smart device usage complies with the power-law appropriation.

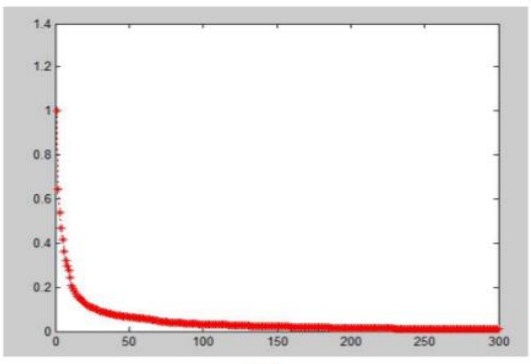
And then the autocorrelation is investigated in this paper. Autocorrelation examination is generally used to mirror the level of correlation between the estimations of a similar arrangement in various time. The initial twenty of autocorrelation are ascertained and appeared in Figure 1(c).



(a) (b) (c)

Fig 2. (a) The cumulative distribution functions of durations. (b) The CDF in log-log coordinates. (c) Autocorrelation of durations.

The last time the device application is interpreted is analyzed. The space here is the sum of two successive start time of the smart device application. This article is filtered at intervals overnight. The average 31.3 minutes, the constant distinction 24.8 minutes, and the coefficient of the variables are 65.8% of all gaps, ie both the biggest and the smallest of the two. CFF, CDF, Lock-Logo Integration and Autocorrelation are shown individually in Figure 2 (a), Figure 2 (B) and Figure 2 (C).



(a)

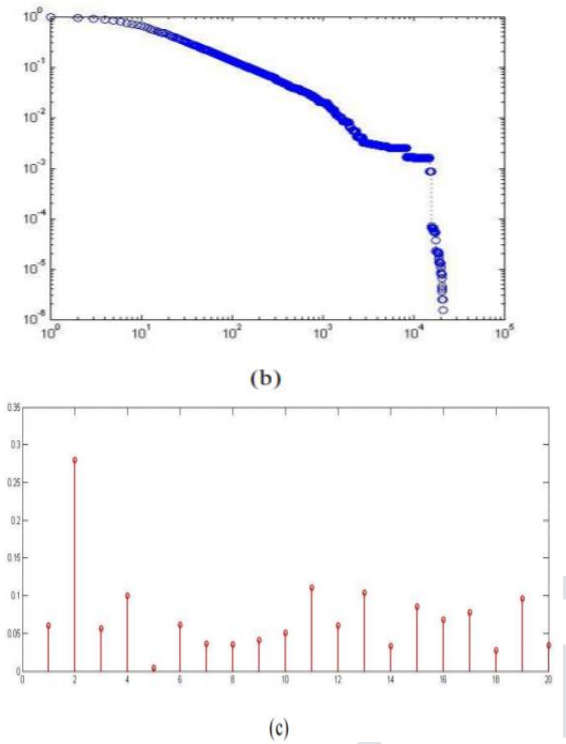


Fig 3: -. (a) The CDF of intervals. (b) The CDF in log-log coordinates. (c) Autocorrelation of intervals.

C. Periodicity of User Behavior

DFT (Discrete Fourier Transform) is regularly used to break down the intermittent behavior has used DFT to understand user behavior. First the idea of dynamic degree is acquainted with measure how dynamic a user is to utilize smart device. For consistently, if a user is utilizing smart device, the dynamic level of this moment is 1, generally 0. For at regular intervals, the dynamic degree is the total of consistently. Ten minutes is the base time unit in the accompanying advances. And then DFT is performed and the PSD (control otherworldly thickness) is appeared in Figure 3.

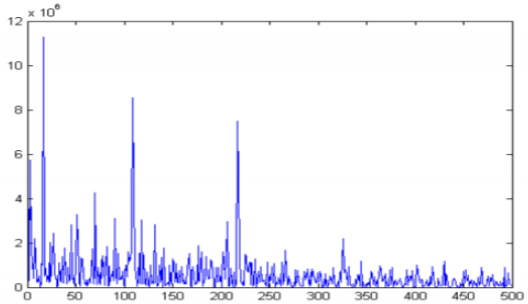


Fig 4: - The power spectral density of user behavior.

In Figure 3, a unit cut is 2 / (NT) = 6.69.010-7 Hz, N = 15645 and T = 10min = 600s. The PSD is a long tail and only 500 values are provided here. In Figure 3, the Business Spectrum Density comes at a peak, whereas the 17th is equivalent to the associated e.g. / (2x3600x17) = equal to 24.4 hours. Therefore, user behavior has been decided that the most obvious time period is 24 hours.

D. Predicting User Behavior

There are not very numerous looks into on predicting user behavior. Of all the existed pertinent works is the most great one. First examined the periodicity of user behavior using DFT, and then used Chebyshev imbalance to predict the best k applications user is most likely to use at a particular time and put these applications on the home screen to make users dispatch their objective applications rapidly. Regardless of the strong hypothetical premise, there is as yet one thing left to be talked about.

That is, user behavior is occasional and the most clear periodicity is 24 hours, as appeared in both and this paper. In any case, when performing predicting, Restrain their concentration in one day and predict the behavior at particular time x utilizing history behavior at other time. For instance, when predicted the behavior at time 15, it utilized the history behavior at time 9:23 and time 22:08.

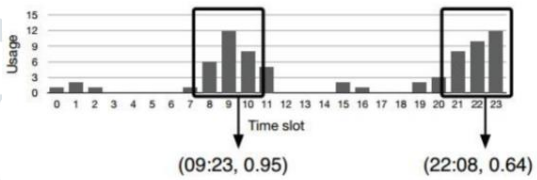


Fig 5: - The predicting approach

There are two unreasonable points in this way. First, the most obvious periodicity is 24 hours, but predicted the behavior using the other time's history behavior. Second, as shown in Figure 1(c) and Figure 2(c), the autocorrelation of duration and interval is neither significant, so to predict user behavior using adjacent behavior is not a good choice.

Here we build a new model to predict user active degree. First let's make clear the problem. The aim is to predict the user behavior at the present day using behavior data from day 1 to day n-1.Let ai,x represent the user active degree at time x in day i. The smallest time unit is ten minutes. Smooth ai, x with ten minutes and turn it into a' i,x. Calculate the mean and variance of ai, x (1 i n-1), and notate as E and V separately.

Use the notation A to represent the real user active degree which is to predict. According to Chebyshev inequality, expression (1) decides the relationship stands for any of value of A and its probability, where P(x) means the probability of event x, positive value.

$$P[\,|A - E| \geq \varepsilon\,] \leq D / \varepsilon^2 \quad (1)$$

$$E_n = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad (2)$$

$$D_n = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - E_n^2 \quad (3)$$

$$E_{n+1} = \frac{x_1 + x_2 + \ldots + x_n + x_{n+1}}{n+1} = \frac{n E_n + x_{n+1}}{n+1} \quad (4)$$

$$D_{n+1} = \frac{x_1^2 + x_2^2 + \ldots + x_n^2 + x_{n+1}^2}{n+1} - E_{n+1}^2 = \frac{n(D_n + E_n^2) + x_{n+1}^2}{n+1} - E_{n+1}^2 \quad (5)$$

E. Methodology

This research need to define the structure of database access log in the form of sequences of actions Ai∈ Ω . Most of database access logs consist of records of similar structure:

(user id, event, sql, time, other features)

where user id is user login; occasion is a kind of occasion (e.g., begin or complete of a query execution); sql will be SQL content of a query; time is a timestamp; different highlights can be divided into execution assemble that incorporates numerical attributes of query execution (e.g. number of read/compose activities, span, and so on); and identification gather with discrete qualities of query, for example, identifiers of customer process, server's procedure, user pseudonyms, and so on.

In this manner the issue is to guide such structure into a limited letters in order. We propose the accompanying technique.

Pre-processing procedure for DB access registration:
Step 1 - Reduce the "inexperienced" attribute
Step 2. - Number attributes descritization
Step 3. - Extracting templates from SQL statement (skeletons)
Step 4 - Unique attributes that combine defined characters.

F. Experiment Result

The objectives of experiment are to check how conventional data mining methods (successive examples and affiliation rules) take a shot at true data with our proposed SQL-follow making an interpretation of strategy and to contrast execution of existing methods with our novel technique, in view of time-subordinate element mapping and choice tree learning calculation. This research thinks about two situations: "next activity expectation" and "inconsistency location".

It run probes genuine data, gathered from MS SQL Server follow logs and created by certifiable saving money intranet application. The undertaking of the application is enrolling, assessing and handling purchaser credit demands. An administrator enters and procedures client's solicitations in the framework.

A few genuine people more often than not work all the while under a similar administrator's login. This gathered hints of administrators' action in one part of the bank amid two days, one day – for preparing, another for testing. There are around 30000 SQL questions for each day. Applying SQL follow change technique This consider just SQL question content, execution time, term and number of read/compose activities in an inquiry.

Table I: - The exploratory execution result

Experiment Settings	*Algorithm*	*Hit Ratio*
Training: 8h (33856 records)	Pf-DT	85.76%
Testing: 8h (28060 records)	Seq-EM	59.72%
No anomalies	A-Rules	42.47%
Training: 4h (16180 records)	Pf-DT	79.77%
Testing: 8h (28060 records)	Seq-EM	43.72%

The principal arrangement of analyses was for "next activity expectation" situation. To examine how the extent of the preparation set influences the model accuracy and arranged three preparing sets of various sizes: 2 hours, 4 hours and 8 hours (the entire working day) of movement. The testing dataset is 8 hours of action in one more day. Preparing time of all calculations in these tests was almost the equivalent, around one moment or less. The exploratory execution results (hit proportion) are introduced in the table beneath:

The second arrangement of tests is given to the examination of the issue of abnormality identification. To evaluate the capacity of the calculations to find abnormalities This have added to the testing dataset 10% of haphazardly created peculiar activities (conceivable activities however in an irregular spots). This attempted 1% and 5% yet the outcomes ended up being fundamentally the same as 10%, that is the reason (and on account of room confinement) This leave results for 10%. They are introduced on ROC bend graph underneath:

Table II: - Comparison table values of False Positive Rate

Existing 1	*Existing 2*	*Proposed*
0.13	*0.09*	*0.02*
0.2	*0.14*	*0.05*
0.28	*0.19*	*0.09*
0.39	*0.25*	*0.14*
0.45	*0.3*	*0.19*

IV. CONCLUSION

Recognizing human activities from sensitive data is a popular research material in the field of mechanical learning. Depending on the receipt of the named data, the authentication methods are simply divided into two categories that are not supervised and supervised.

The proposed large-scale analyzes that are closer to the truth. However, the proposed structure suffers from many limitations. Similar activities are mobilized together and are hard to distinguish. To prevent this, a better way to represent categories or additional features (eg time and time, the duration of the operation) can be defined.

REFERENCES

[1]. Amirah Mohamed Shahiri, Wahidah Husain, Nuraini Abdul Rashid. A Review on Predicting Student's Performance Using Data Mining Techniques. Elsevier, Procedia Computer Science, Volume 72, 2015. Pages 414-422.

[2]. Amjad Abu Saa. Educational Data Mining & Students' Performance Prediction.International Journal of Advanced Computer Science and Applications, Volume 7, No. 5, 2016.

[3]. Brijesh Kumar Bhardwaj, Saurabh Pal. Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security, Volume 9, N0. 4, April 2011.

[4]. Zachary A. Pardos, Neil T. Hefferman, Brigham Anderson, and Cristina L. Heffernan.The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks. Springer- Verlag Berlin Heidelberg 2007, pages 435-439.

[5]. EdinOsmanbegovic, MirzaSuljic. Data Mining Approach For Predicting Student Performance, Economic Review- Journal of Economics and Business, Volume X, Issue 1, May 2012.

[6]. GurmeetKaur, Williamjit Singh. Prediction of Student Performance Using Weka Tool, Research Cell: An International Journal of Engineering Sciences, January 2016, Volume 17.

[7]. SuchitaBorkar , K. Rajeswari.Predicting Students Academic Performance Using Education Data Mining, International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 7, July 2013, pages 273 – 279.

[8]. Shiwani Rana, Roopali Garg. Student's Performance Evaluation of an Institute Using Various Classification Algorithms Information and Communication Technology for Sustainable Developmet, November 2017, pages 229-238.

[9]. Vignesh Ramamoorthy.H and Balakumaran.P.J, 'Evolving an E-Governance System for Local Self-Government Institutions for Transparency and Accountability' published in International Journal of Information Engineering and Electronic Business (IJIEEB), Volume 5, No.6, December 2013, Page – 40 to 46, ISSN: 2074-9023 (print), ISSN: 2074-9031 (online).

[10]. Vignesh Ramamoorthy.H, 'An Encrypted Technique with Association Rule Mining in Cloud Environment' published in International Journal of Computer Applications (IJCA), Foundation of Computer Science, New York, USA, 2012, Page – 5 to 8, ISBN: 973-93-80867-88-1.

[11]. Vignesh Ramamoorthy.H, 'Bigdata Analytics: Comparative Study of Tools' published in International Journal of Computer Science (IJCS), Volume 5, Issue 1, No 2, 2017, Page – 995 to 1003, ISSN: 2348-6600.

[12]. Vignesh Ramamoorthy.H, 'An Analysis on Indexing Techniques for Scalable Record Linkage, Data Leakage and De-duplication in World Wide Web' published in International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Volume 4, Issue 3, March 2016, Page – 3150 to 3156, ISSN(Online): 2320-9801, ISSN (Print) : 2320-9798

[13]. E.I. Georga, V.C. Protopappas, D. Polyzos, D.I. FotiadisEvaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models Med Biol Eng Comput, 53 (12) (Dec 2015), pp. 1305-1318, 10.1007/s11517–015–1263-1[Epub 2015 Mar 15] CrossRefView Record in ScopusGoogle Scholar

[14]. B.J. Lee, J.Y. Kim, B.J. Lee, J.Y. KimIdentification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning IEEE J Biomed Health Inform, 20 (1) (Jan 2016), pp. 39-46, 10.1109/JBHI.2015.2396520 [Epub 2015 Feb 6] CrossRefView Record in ScopusGoogle Scholar

[15]. C.R. Marling, N.W. Struble, R.C. Bunescu, J.H. Shubrook, F.L. SchwartzA consensus perceived glycemic variability metric J Diabetes Sci Technol, 7 (4) (Jul 1 2013), pp. 871-879 CrossRefView Record in ScopusGoogle Scholar

[16]. J.H. Huang, R.H. He, L.Z. Yi, H.L. Xie, D.S. Cao, Y.Z. LiangExploring the relationship between 5′AMP-activated protein kinase and markers related to type 2 diabetes mellitus Talanta, 110 (Jun 15 2013), pp. 1-7, 10.1016/ j.talanta.2013.03.039 [Epub 2013 Mar 22] ArticleDownload PDFCrossRefView Record in ScopusGoogle Scholar

[17]. A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul Quantitative population–health relationship (QPHR) for assessing metabolic syndrome EXCLI J, 12 (Jun 26 2013), pp. 569-583 [eCollection 2013] View Record in ScopusGoogle Scholar M.W. Aslam, Z. Zhu, A.K. NandiFeature generation using genetic programming with comparative partner selection for diabetes classification Expert Syst Appl, 40 (13) (2013), pp. 5402-5412 ArticleDownload PDFView Record in ScopusGoogle Scholar

[18]. C. Sideris, M. Pourhomayoun, H. Kalantarian, M. SarrafzadehA flexible data-driven comorbidity feature extraction framework Comput Biol Med, 73 (Jun 1 2016), pp. 165-172, 10.1016/j.compbiomed.2016.04.014 [Epub 2016 Apr 20] ArticleDownload PDFView Record in ScopusGoogle Scholar

[19]. L. BreimanRandom forests Mach Learn, 45 (1) (2001), pp. 5-32, 10.1023/A:1010933404324 CrossRefView Record in ScopusGoogle Scholar M. Robnik-Sikonja, I. KononenkoTheoretical and empirical analysis of ReliefF and RReliefF Mach Learn, 53 (1–2) (2003), pp. 23-69, 10.1023/ A:1025667309714 CrossefView Record in ScopusGoogle Scholar

[20]. T.M. Cover, P.E. HartNearest neighbor pattern classification IEEE Trans Inf Theory, IT-13 (1) (1967), pp. 21-27 CrossRefView Record in ScopusGoogle Scholar.

[21]. L.F. Chen, C.T. Su, K.H. ChenAn improved particle swarm optimization for feature selection Intell Data Anal, 16 (2) (2012), pp. 167-182 CrossRefView Record in ScopusGoogle Scholar

[22]. P.Sathya "An Investigation on Predictive Pattern Mining and Temporal Data Features Extraction in Data Mining" in Volume 7, Issue 4(July, 2018) of our journal IJSRCSAMS