AIJREAS VOLUME 1, ISSUE 4 (2016, April) (ISSN-2455-6300) Online
Anveshana's International Journal of Research in Engineering and Applied Sciences

Investigation on Association Rules Mining Based Scheme for Healthcare Repository

D.RAVI KIRAN
Research Scholar, Acharya Nagarjuna University,
Nagarjuna Nagar, Guntur(D.t)
Email: ravikirand2005@gmail.com

Dr. S V N SRINIVASU
Research supervisor, Acharya Nagarjuna University,
Nagarjuna Nagar, Guntur(D.t)

Abstract:

Early detection of patients with elevated risk of developing various disease is a challenging issue, in this connection Healthcare institutes enrich the repository of patients' disease related information in an increasing manner which could have been more useful by carrying out relational analysis. Data mining algorithms are proven to be quite useful in exploring useful correlations from larger data repositories. In this paper we have investigation on Association Rules mining based scheme for finding co-occurrences of diseases carried by a patient using the healthcare repository. In this paper we proposed prediction of clinical outcome (PCO) scheme which extracts data from patients' healthcare database, transforms the OLTP data into a Data Warehouse by generating association rules. The PCO system helps reveal relations among the diseases. The PCO system predicts the correlation(s) among primary disease and secondary disease/s .

Key words *Disease Correlation, Association Mining, e-Health, Healthcare, Medicare*

1. Introduction

Now a day's medical science has exposed that the frequency of one disease can lead to several associated diseases [1]. For example, Heart-Block can lead to the occurrences of other diseases like Hypertension, Cardiac Arrest and so on. It is, however, still an interesting problem [1], to see how far the medical philosophy holds from statistical point of view. Data mining based techniques, like association rule mining, have gained popularity [1-3] among contemporary scientists to gain clearer understanding of different physical and scientific phenomenon. In this paper, we apply association rule mining to extract knowledge from clinical data for predicting correlation of diseases carried by a patient. From the viewpoint of scientific research, data mining is relatively a new discipline that has been developed mainly from studies carried out in various disciplines such as computing, marketing, statistics and so on [4]. Data mining problems and corresponding solutions have roots in classical data analysis. Many of the methodologies used in data mining has come from two branches of research: i) one is developed in machine leaning (artificial intelligence) community, and ii) the other is developed in the statistical community particularly in multivariate and computational statistics. Both have made great contributions to the understanding and applications of data mining techniques [5, 6]

2. Back ground

Computerized healthcare systems are accumulating large quantities of information about patients and their medical conditions every day. Unfortunately, few methodologies have been developed and applied to discover this hidden knowledge [7]. The cluster-analysis based model is suggested and discussed [8] for assigning prostate cancer patients into homogenous groups with the aim to support future clinical treatment decisions as an illustration. To explore association rules in noisy and high dimensional medical data-repository an improved algorithm has been introduced with several constraints [9]. A statistical analysis of decision tree based classification approach on diagnosing the Ovarian Cancer using Bio-marker Patterns Software (BPS) has been applied [10]. A task [11] has been accomplished on comparison of data mining methods supporting diagnosis for Melanoma. Association

rule classifiers have been applied to diagnose breast cancer using digital mammograms [12], Neural Network based classification approach also used for the same purpose [13]. Association Mining applied on questionnaire responses related to human sleeping [14] where questionnaire data and clinical summaries comprised a total of 63 variables including gender, age, body mass index, and Epworth and depression scores. Many Clinical Decision Support Systems (CDSS) have been developed. CHICA [15] is a CDSS, developed to improve preventive pediatric primary care. Dynamic forms are generated and tailored to patients' needs based on the Medical Logic Modules (MLMs). A knowledge management framework for distributed healthcare systems has been proposed [16] to integrate the heterogeneous systems used by different departments from clinical care to administration. However, the aforementioned developments are application specific and thus hard to apply in general. Instead of developing an application limited to a specific purpose such as prostate cancer [8], skin cancer [11], and sleeping [14] and so on, we proposed for a more generic version of PCO system that can work for all diseases in similar fashion and generate correlations depending on the input dataset. In the next section, the architecture of the application and its working procedures are stated.

3. Framework and Working Procedures

The PCO system will extract data from an OLTP system. Thus to implement the PCO system, we require the regular operational OLTP system to generate input data for the PCO system. The architecture of the system can be described as follows.

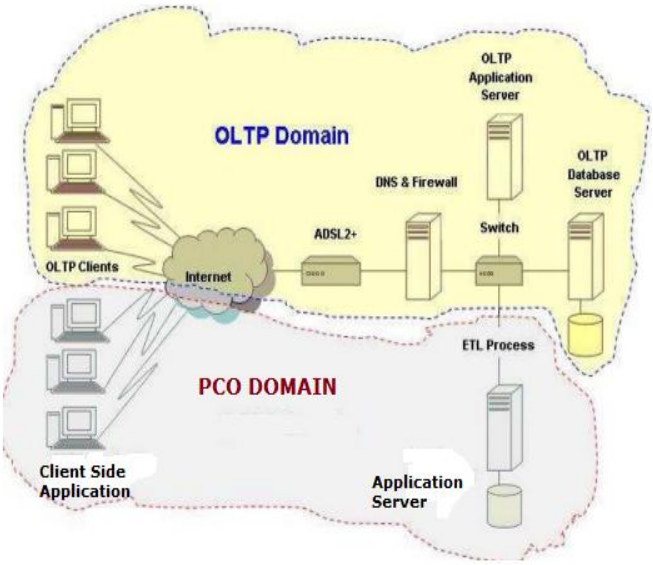

Figure 1 - The Complete Architecture of the Proposed System for Prediction Correlation among diseases using association mining on healthcare data. The upper-portion of the figure describes the architecture of the OLTP system and the lower portion is the PCO system.

3.1 System Architecture

The disease diagnosis application proposed here based on two different software systems. One is online transaction processing (OLTP) system and another is prediction of clinical outcome (PCO) application. **OLTP Application:** When a patient will visit a doctor he/she must have to fill up a prescribed form and the information from the form is proposed to be captured through web enabled OLTP system.

PCO Application: This application will import transactional records from OLTP application for further processing to generate correlations among diseases using Association Rules data mining. The OLTP system is a regular database application to capture patient's information and to preserve the records into a database repository. It is simple and no analytical process has been incorporated in this portion and thus the PCO system is our major focus in this paper and we presume that the OLTP system has been

implemented successfully and data has been captured accordingly from healthcare institutes.

3.2. Association Rules Mining and Apriori Algorithm

In data mining, association rule is used for discovering interesting relations between variables in large databases. The two key terms support and confidence are used in computing correlations between variables which can be defined as follows:

Support: In a fixed number of transactions the occurrences of a particular event is the support of that event. For example there are T transactions among which Txy transaction s contain the itemset {X,Y} is Tx.

Confidence: confidence is a relative support. For example there are T transactions among which Txy transactions contain itemset {X, Y} and Tx transactions contain item X and thus confidence of occurring X and Y together is Txy/ Tx. The Apriori Algorithm [17] is used here for Association Rule Mining to find out frequent dataset that satisfy the predefined minimum support and confidence from a given database. As computers are handy now-a-days, a viable PCO system can be setup to collect large volumes of data and to stored in the database simultaneously. This kind of data includes the transaction records of clinics, hospitals, supermarkets, banks, stock markets, telephone companies and so on. The next few sections we try to discover some hidden information from a sample transactional dataset.

3.2 Sample Dataset

In a clinic various patients come but most of them come for a particular disease, which mentioned here as primary disease. When Doctors or their associates make an interview with the patient and note down other problems (diseases), which mentioned here as secondary or associated diseases are inserted into a database. The main objective of the paper is to find out relations among the primary disease and other secondary diseases. The following table represents sample dataset of a Medicare database that contains the patient-wise diseases.

Table 1. - Transactional data sample

Patient Id	Disease
P000000001	Bradycardia
P000000001	Cardiac Arrest
P000000001	Hypertension
P000000001	Myocarditis
P000000002	Bradycardia
P000000002	Cardiac Arrest
P000000002	Hypertension
…	…
P000001000	Cardiac Arrest

Table 2: Patient records with multiple diseases

Patient	Count	Diseases
P000000001	4	Heart-Block, Hypertension, Cardiac-Arrest, Bradycardia
P000000002	3	Heart-Block, Hypertension, Cardiac-Arrest
…	…	…
P000001000	1	Hypertension

3.3 Producing Itemsets

A set of diseases obtained by each patient presented here along with the number of diseases counted by the algorithm

Algorithm 1 - CountDisease

```
1   PROCEDURE CountDisease
2     FOR each p in P
3         D_s <= ""
4         c <= 0
5         find records for p
6         FOR each r in R
7             c <= c+1
8             D_s <= D_s + r + ","
9         NEXT r
10        D_s <= D_s without comma at the end
11        INSERT (p, c, D_s) in to database table
12    NEXT p
13  END CountDisease
```

Figure 2 - Pseudo code to count diseases carried by any patient.

3.4 Counting Support of an Item (Disease) in the sample dataset

The frequency of every item in all the transactions has been calculated in the following table implementing the following algorithm and for a 1-item itemsets for the first pass.

```
    Algorithm 2- FindSupport
1   PROCEDURE FindSupport
2     d ⇐ disease
3     db ⇐ database in considarat ion
4     r ⇐ record
5     rs ⇐ recordset
6     s ⇐ support
7     BEGIN
8           FOR each d in db
9             s ⇐ 0
10        find records for d
11            FOR each r in rs
12              s ⇐ s + 1
13            NEXT r
14        UPDATE Database with s
15          NEXT d
16    END
17  END FindSupport
```

Figure 3 - Pseudo code to calculate support.

3.5 Association Rule Generation

Using the sample transactions of Table 1, after second pass (maxpass=2), the PCO system has generated the rule data as listed in Table 3 below

Table 3- Candidate Sets for stop-level 2

Itemset	Supp (%)	Conf (%)
{Bradycardia, Cardiac-Arrest}	2.60	8.52
{Bradycardia, Heart-Block}	0.90	2.69
……	….	…..
{Heart Block, Hypertension}	1.00	1.82
{Heart Block, Myocarditis}	1.10	3.29

4. Conclusions

In this paper, I have applied system-prototype, named PCO system, using the association rules of data mining technique applied to a patients' (assumed) database for discovering patterns of diseases that might be carried by a patients. As a novel idea of mining the data capturing process can further be modified in the clinics as well as in the data-warehouses which should further be involved to enhance the PCO system we have proposed. The recognized pattern by this implementation definitely can improve the healthcare services along with medical researchers for further exploring trends of diseases that are correlated. To make sure strapping national economy and bio-security.

References

[1] Lovell NH, Magrabi F, Celler BG, Huynh K, and Garsden H, Web-based acquisition, storage, and retrieval of biomedical signals, IEEE Eng. Medicine and Biology, vol. 20(3), pp. 38-44, 2001.

[2] Mazzi C, Ganguly P, and Kidd M, Healthcare application based on software agents, in Medinfo 2001 Proceedings, 2001, pp. 136-140.

[3] Takeuchi H, Kodama N, Hashiguchi T, and Hayashi D, Automated Healthcare Data Mining Based on a Personal Dynamic Healthcare System,

the 28th IEEE EMBS Annual International Conference, New York City, USA, 2006.

[4] Giudici P, Applied Data Mining: Statistical Methods for Business and Industry: Wiley Publications, 2003.

[5] Mannila H, Data mining: machine learning, statistics, and databases, in Eighth International Conference on Scientific and Statistical Database Management, 1996, pp. 1-8.

[6] Kantardzic M, Data Mining: Concepts, Models, and Algorithms. NJ: Wiley Interscience, 2003.

[7] Prather J, Lobach D, Goodwin L, Hales J, Hage M, and Hammond W, Medical data mining: knowledge discovery in a clinical data warehouse, in Proc AMIA Annu Fall Symp., 1997, pp. 101-5.

[8] Churilov L, Bagirov A, Schwartz D, Smith K, and Dally M, Improving Risk Grouping Rules for Prostate Cancer Patients with Optimization, the 37th International Conference on System Sciences, Hawaii 2004. [9] Ordonez C, Santana C, and Braal LD, Discovering Interesting Association Rules in Medical Data, in ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2000, pp. 78-85. [10]Vlahou A, Schorge JO, Gregory BW, and. Coleman RL, Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data, Journal of Biomedicine and Biotech-nology, vol. 5, pp. 308– 314, 2003.

[11]Grzymala-Busse JW and Hippe JS, Data Mining Methods Supporting Diagnosis of Melanoma, the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS 2005), 2005, pp. 371–373.

[12]Za¨ıane OR, Antonie ML, and Coman A, Mammography classification by an association rulebased classifier, in MDM/KDD: Inter-national Workshop on Multimedia Data Mining (with ACM SIGKDD 2002), 2002, pp. 62–69.

[13]Land WH, Masters JT, Lo JY, McKee DW, and Anderson FR, New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data in Industrial Applications, Proceedings of the 2001 IEEE Mountain Workshop on Soft Computing, Virginia Tech, Blacksburg. Virginia, 2001, pp. 47–52.

[14]Laxminarayan P, Ruiz C, Alvarez SA, and Moonis M, Mining associations over human sleep time series, the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS), 2005, pp. 323–328. [15]Anand A, Biondich PG, Liu G, Rosenman M, and Downs SM, Child health improvement through computer automation: The CHICA system, in MEDINFO, 2004.

[16]Kazemzadeh RS and Sartipi K, Interoperability of Data and Knowledge in Distributed Health care Systems, in The 13th IEEE International Workshop on Software Technology and Engineering Practice, 2005.

[17]Agrawal and Srikant, Fast algorithm for Mining Association Rules, in The 20th VLDB conference Santiago, Chile, 1994.

[18]A Healthier Future For All Australians. vol. ISBN: 1- 74186-940-4 Publications Number: P3 - 5499: Commonwealth of Australia, 2009. [19]Medicare Australia, www.medicareaustralia.gov.au, last access: Oct, 2009.