Proposal for Enhancing Medical Diagnosis of Disease Related With Patients Environment

Zahraa A. Saed

University of technology, computer science department

Abstract

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships that can be hidden among vast amount of data. This research introduces a proposal to improve and enhance a medical diagnosis using association rules of data mining technique, especially for medical diagnosis of diseases related with patient's environment. That proposal found new relationships and predications to support early medical diagnosis, that by build a two proposed databases: the first data base contained a basic attributes of blood and tissues for the patients. The second database contained a basic attributes of patients profile and environment. After finding all associations rules from these two proposed databases, these rules will be mixed by a proposed method to gain a new rules give new patterns will predict relations among the physiology and environment of patients and disease.

Keywords: data mining, association rules, medical diagnosis and environment disease

المستخلص

استخراج البيانات هي العملية التي تستخدم مجموعة متنوعة من أدوات تحليل البيانات لاكتشاف الأنماط والعلاقات التي يمكن أن تكون مخفية بين كمية هائلة من البيانات. هذا البحث يقدم اقتراحا لتحسين وتعزيز التشخيص الطبي باستخدام قواعد رابطة تقنية التنقيب عن البيانات، وخاصة في مجال التشخيص الطبي من الأمراض المرتبطة مع بيئة المريض. وجدت أن الاقتراح علاقات جديدة والتنبؤات لدعم التشخيص الطبي المبكر، وذلك عن طريق بناء قاعدتي بيانات مقترحة: قاعدة البيانات الأولى تحتوي على سمات أساسية من الدم والأنسجة للمرضى. قاعدة البيانات الثانية تحتوي سمات الشخصية الأساسية للمرضى والبيئة. وبعد العثور على جميع القواعد المترابطة من هاتين القاعدتين المقترحتين، سيتم مزج هذه القواعد من خلال طريقة مقترحة للحصول على قواعد جديدة تعطي أنماط جديدة للتنبؤ بالعلاقات بين بيئة المرضى وفسلجة اجسامهم والمرض.

1. Introduction

Data mining derives its name from the similarities between searching or valuable information in a large database and mining rocks for a vein of valuable ore. The more general terms such as Knowledge Discovery in Databases (KDD) describe a more complete process. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. The efficient discovery of such rules has been a major focus in the data mining research community. Many algorithms and approaches have been proposed

to deal with the discovery of different types of Association Rules (AR) discovered from a variety of databases [1, 2].

The problem is stated as follows, Let $I = \{i1, i2, ...im\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. A unique identifier *TID* is given to each transaction. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ has a *support s* in the transaction set D is s% of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if c% of transactions in D that contain X also contain Y . In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater that given thresholds. These rules are called *strong rules* [3, 4].

Mining of association rules from a database consists of finding all rules that meet the user-specified threshold support and confidence. The problem of mining association rules can be decomposed into two subproblems as stated in Algorithm 1 [5, 6].

Algorithm 1. Basics

Input: I (Itemset), D(Database), s(support), c(confidance)

Output: Association rules satisfying s and c

Process:

1) Find all sets of items which occur with a frequency that is greater than or equal to the user-specified threshold support, s.

2) Generate the desired rules using the large itemsets, which have user-specified threshold confidence c.

The first step in Algorithm 1 finds *large* or *frequent itemsets.* Here an itemset is a subset of the total set of items of interest from the database. An interesting (and useful) observation about large itemsets is that:

If an itemset X is small, any superset of X is also small.

Of course the contrapositive of this statement (If X is a large itemset then any subset of X is also large) is also important to remember. In the remainder part of this chapter L is used to designate the set of large itemsets. The second step in Algorithm 1 finds association rules using large itemsets obtained in the first step. The identification of the large itemsets is computationally expensive. However, once all sets of large itemsets ($l \in$ L) are obtained, there is a straightforward algorithm for finding association rules which is restated in Algorithm 2 [5, 6].

Algorithm 2. Find Association Rules Given Large Itemsets:

Input: I(Itemset), D(Database), s(support), c(confidence), L(Large itemset)

Output: Association rules satisfying s and c

Algorithm:

1) Find all nonempty subsets, *x*, of each large itemset, *l* ∈ L

2) For every subset, obtain a rule of the form x⇒ (*l*-x) if the ratio of the frequency of occurrence of *l* to that of x is greater than or equal to the threshold confidence.

2. The Proposed System

The proposed system aim to enhance medical diagnosis especially for disease related with patient environment. That will be done by a proposed system, to explain the system in detail follow the consequence levels:

2.1 Build the two proposed databases

Build two proposed database, first database called physiology database which deals with human critical elements in medical diagnosis. So it will have the following attributes:

Red Blood Cells (RBC), no. of RBC, White Blood Cells (WBC), no. of WBC, no. of cells, size of cells, nuclease shape, and rate of division.

The suggested encoding of these attributes as in the following:

1. **RBC: if it was in normal range then A will appear, but if not, A not will appear.**
2. **No. of RBC: if it was in normal range then B will appear, but if not, B will not appear.**
3. **WBC: if it was in normal range then C will appear, but if not, C will not appear.**
4. **No. of WBC: if it was in normal range then D will appear, but if not, then D will not appear.**
5. **No. of cells: if it was in normal range then E will appear, but if not, E will not appear.**

6. **Size of cells: if it was in normal range then F will appear, but if not, F will not appear.**
7. **Nuclease shape: if it was in normal range then G will appear, but if not, G will not appear.**
8. **Rate of division: if it was in normal range then H will appear, but if not, H will not appear.**

See figure (1) which show the proposed encoded physiology database.

TID	RBC	no. of RBC	WBC	no. of WBC	no. of cells	size of cells	nuclease shape	Rate of division
1	A		C	D		F	G	H
2	A	B	C	D	E			
.			C	D	E	F	G	H
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	..		.		

Figure (1): The proposed encoded physiology database for medical diagnosis.

Second database called environment database which deal with human social specifications, environment, so it will have the following attributes:

Age, sex, marriage state, children no. salary, education level, smoking, drinking and inherent diseases. The suggested encoding of these attributes as in the following:

1. **Age: if it was above 30 years I will appear, but if not, I will not appear.**
2. **Gender: if it was female J will appear, but if not, J will not appear.**
3. **Marriage state: if it was marriage K will appear, but if not, K will not appear.**

4. **Children no.: if it was above 3 L will appear, but if not, L will not appear.**
5. **Salary: if it was above reasonable range M will appear, but if not, M will not appear.**
6. **Education level: if it was above reasonable level N will appear, but if not, N will not appear.**
7. **Smoking: if was do O will appear, but if not, O will not appear.**
8. **Drinking: if was drinking P will appear, but if not, P will not appear.**
9. **Inherent diseases: if was has Q will appear, but if not, Q will not appear.**

See figure (2) which show the proposed encoded environment database.

TID	Age	Sex	mariage state	children no.	Salary	education level	Smoking	drinking	inherent diseases
1	I	J	K	L				P	Q
2			K	L	M	N	O		
.	I	J	K	L	M	N	O	P	Q
.									
.									
.									

Figure (2): The proposed encoded environment database for medical diagnosis.

2.2 The proposed Mixer

First extracting the association rules for the two proposed databases then will mix these association rules by using Proposed Method to obtain new rules for new predictions, to explain the proposed method in details will introduce the following proposed algorithm:

Proposed Algorithm for Mixing

Input: ***The two proposed encoded databases (the displayed in fig. (1) and fig. (2))***

Output: mixed association rules

Process:

Step1: extracting association rules from proposed physiology database with minimum support = 50% and minimum confidence = 50%.

Step2: extracting association rules from proposed environment database with minimum support = 50% and minimum confidence = 50%.

Step3: mixing the extracted association rules by the following steps:

1. **Make File1 to be the file has association rules extracted from the physiology database.**
2. **Make File2 to be the file has association rules extracted from the environment database.**
3. **While not end of File1**
    - **Take the current association rule (AR1)**
    - **Mixing it with all association rules in File2 (AR2) by the form AR1---->AR2**
    - **The minimum confidence of the new mixed rules proposed to = 60%, the confidence of this new rule will be calculated as arrange of the two confidence of both AR1 and AR2.**
    - **If the confidence of mixed rules equal or greater than 60% this mixed rules will be inserted in File3 which store the mixed rules, else will be omitted.**
4. **Analyzing (display encoded AR to it is original attributes) all rules in File1, File2 and File3.**

5. **Visualize the new relations and pattern for prediction and enhancing diagnosis the disease related not just with physiology but also with environment.**
6. **End.**

3. The Implementation and Discussion

To explain the proposed method practically, introduce the proposed program, see figure (3) which explain the main window for the program. This window has three commands these are: mining physiology database, mining environment database and mixing physiology AR and environment AR.

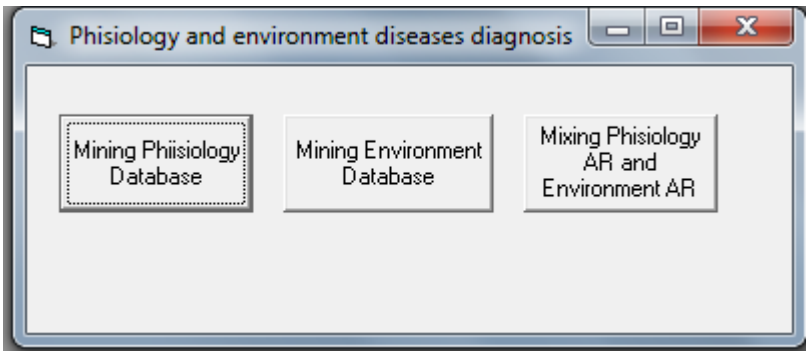

Figure (3): main window of implementation.

When user clicked the first command, mining physiology database, in figure (3), then the program will connect with the proposed encoded physiology database and applying association rules mining on it and finally display the window in figure (4), which display minimum support, minimum confidence and file name the association rules was stored in it.

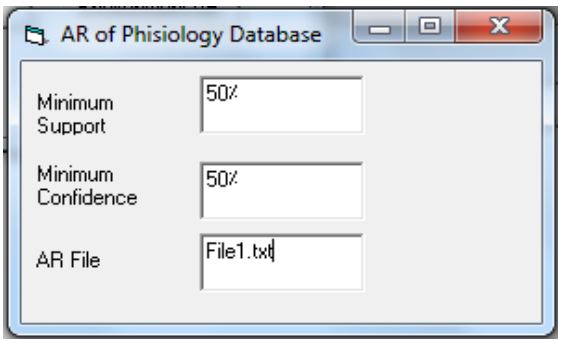

Figure (4): AR information of physiology database

When user clicked the second command, mining environment database, in figure (3), then the program will connect with the proposed encoded environment database and applying association rules mining on it and finally display the window in figure (5), which display minimum support, minimum confidence and file name the association rules was stored in it.

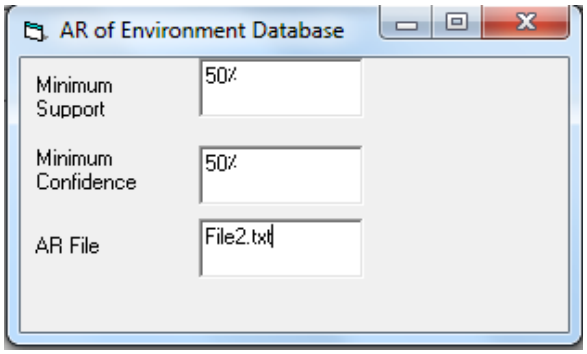

Figure (5): AR information of environment database

When user clicked the third command, mixing physiology AR and environment AR, in figure (3), then the program will connect window in figure (6) which require from the user the files names to be mixed according the proposed method, and after activate the command mixing the software will apply mixing method and store the mixed rules in new file called Fil3.txt, see figure (7) which display that file.

Figure (6): AR mixing method

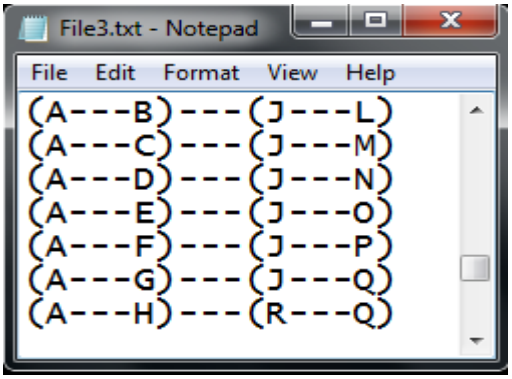

Figure (7): mixed AR file

Here get a new association rules which give us new relations among patient's physiological, patient's environment and disease. These relations could give a new prediction to diagnosis diseases in very early times. From these relations will introduce the following, after analysis stage there are new rules which predicate very strong relations, these relations compared with most new reports in USA with HIV patients [7], these are:

1. **Prevalence is the number of people living with HIV infection at the end of a given year. At the end of 2011, an estimated 2,106,400 persons in the United States were living with HIV infection has age above30, in the proposed program see the drinking and smoking make change in this estimated rate, since drink and smoke modify blood component attributes, so the program find with 21% of patients are undiagnosed.**
2. **Incidence is the number of new HIV infections that occur during a given year. In 2011, estimation approximately 99,300 people were newly infected with HIV those people Black/African American men, in the proposed program find**

this estimation rate are incrassated if they have also low salary.

3. Of the estimated number of diagnoses of HIV infection in the USA states with confidential name-based HIV infection reporting in 2011.

The distribution of ages and no. of patient's diagnosis was as in table (1), the depended sample was as in USA 2011 report (2 million patients [11]):

Table (1): infection HIV estimation with 2011 reports and proposed program.

Age	No. of HIV infections diagnosis (proposal)	No. of HIV infections diagnosis (USA 2011 report)
Under 13	200	150
Ages 13-14	30	30
Ages 15-19	3,778	3,433
Ages 20-24	8,390	8,200
Ages 25-29	7,994	7,194
Ages 30-34	7,009	6,809
Ages 35-39	7,433	7,234

4. Conclusions

From the implementing the proposed mixing association rules, conclude the following:

1. Since the proposal deal with disease related with patient physiology and environment, research introduces to proposed databases for sample of patients. These are physiology database which contain some most critical attributes of blood and tissues related with diseases diagnosis and environment database which contain some most critical attributes of patient environment.
2. Each database has 1000 transaction (each one has the medical information for the same 1000 person). Those persons are the

1000 patients are varying from healthy to suspicious to early infected and finally to infected patients. That for obtaining more various patterns and relationships between various patients under the umbrella of these diseases related with environment.

3. **Applying proposal for building two proposed databases instead of mixing all attributes for physiology and environment in one database safe the time and space in extracting association rules.**
4. **proposing mixing method for mixing the extracted association rules in both databases give us more suitable and arranged pattern so it be more easily for analysis since it will always in the following shape AR1--->AR2.**
5. **We customize the mixing algorithm for proposed technique, that by propose a schema for encoding the rules, and then making the left part of mixed rules is physiology rules and right part is pool is environment rules.**
6. **Mixing confidence customized to satisfy the basic conditions in building the association rules according the data mining techniques.**

References

1. **M. S. Chen, J. Han, and P. S. Yu. "Data mining: An overview from a database perspective". IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.**
2. **J. Han and M. Kamber. "Data Mining: Concepts and Techniques". Morgan Kaufmann, 2000.**
3. **Mitra S., and Ahharya T., "Data Mining Multimedia, Soft Computing, and Bioinformatics", John Wiley and Sons, Inc., 2003.**
4. **Uno T., Kiyomi M., and Arimura H., "*LCM ver.2: Efficient Mining Algorithms for Frequent, Closed, Maximal Itemsets*" Japan, 2004.**

http://research.nii.ac.jp/~uno/papers/0411/cm2.pdf

5. **El-Hajj M. and Zaiane O. R., "*YAFIMA: Yet Another Frequent Itemset Mining Algorithm*", Department of Computer Science, University of Alberta, Edmonton, AB, Canada, 2005.**
   http://WWW.cs.ualberta.ca/~zaiane/postscript/idm05.pdf
6. **Moonesinghe H. D. K., Foda S., and Tan P. N., "*Frequent Closed Itemset Mining Using Prefix Graphs with an Efficient Flow-Based Pruning Strategy*", Department of Computer Science and Engineering, Michigan State University,2005.**
7. **http://www.usgovernmentbenefits.org/hd/index.php?t=hiv+statistics+2011**