AN OVERVIEW : PREDICTING MEDICAL DISEASES USING DATA MINING

Mrs. Shweta A.Gode[1], Dr. G.R.Bamnote[2]

Abstract: ***Existing research in association with mining has focused mainly on how to expedite the search for frequently co-occurring groups of symptoms in "medical diseases" type of transactions; less attention has been paid to methods that exploit these "frequent symptoms list" for prediction purposes. This project contributes to the latter task by proposing a technique that uses partial information about the contents of a medical diseases for the prediction of what else the physician is likely to diagnose. Using the recently proposed data structure of item set trees (IT-trees), we obtain, in a computationally efficient manner, all rules whose antecedents contain at least one symptom from the incomplete disease. This project combine these rules by uncertainty processing techniques, including the classical Bayesian decision theory and a new algorithm based on the Dempster-Shafer (DS) theory of evidence combination.***

Keywords:Disease, symptoms, association mining, DS theory.

1. INTRODUCTION

The primary task of association mining is to detect frequently co-occurring groups of items in transactional databases. The intention is to use this knowledge for prediction purposes: if bread, butter, and milk often appear in the same transactions, then the presence of butter and milk in a shopping cart suggests that the customer may also buy bread. More generally, knowing which items a shopping cart contains, we want to predict other items that the customer is likely to add before proceeding to the checkout counter.

This paradigm can be exploited in diverse applications. For example, in the domain discussed in each "shopping cart" contained a set of hyperlinks pointing to a Web page in medical applications, the shopping cart may contain a patient's symptoms, results of lab tests, and diagnoses; in a financial domain, the cart may contain companies held in the same portfolio; and Bollmann- Sedorra et al proposed a framework that employs frequent item sets in the field of information retrieval [2]. In all these databases, prediction of unknown items can play a very important role.

For instance, a patient's symptoms are rarely due to a single cause; two or more diseases usually conspire to make the person sick. Having identified one, the physician tends to focus on how to treat this single disorder, ignoring others that can meanwhile deteriorate the patient's condition. Such unintentional neglect can be prevented by subjecting the patient to all possible lab tests. However, the number of tests can undergo is limited by such practical factors as time, costs, and the patient's discomfort. A decision-support system advising a medical doctor about which other diseases may accompany the ones already diagnosed can help in the selection of the most relevant additional tests. The prediction task was mentioned early as in the

pioneering association mining paper by Agrawal et al., but the problem is yet to be investigated in the depth it deserves. In our work, we wanted to make the next logical step by allowing any symptom to be treated as a class label its value is to be predicted based on the presence or absence of other symptoms. Put another way, knowing a subset of the disease symptoms, we want to "guess" (predict) the rest. In our work, we sought to solve both of these problems by developing a technique that answers user's queries in a way that is acceptable not only in terms of accuracy, but also in terms of time and space complexity.

2. EXISTING SYSTEM

Existing research in association mining has focused mainly on how to expedite the search for frequently co-occurring groups of items in "shopping cart" type of transactions; less attention has been paid to methods that exploit these "frequent item sets" for prediction purposes.

Existing system Disadvantages:

- They didn't find missing items in frequently used item set.
- Couldn't find number of users per item set.
- Time complexity
- Lack of viewing items to the user.

3. LITERATURE SURVEY

The literature survey in indicates that most authors have focused on methods to expedite the search for frequent itemsets, while others have investigated such special aspects as the search for time-varying associations or the identification of localized patterns.

Association mining systems that have been developed with classification purposes in mind are sometimes dubbed classification rule mining. Some of these techniques can be adapted to our needs. Take, for instance, the approach proposed in [12]. If ij is the item whose absence or presence is to be predicted, the technique can be used to generate all rules that have the form r(a) => Ij where , r(a) C (I\i(j) is the binary class label (ij= present or ij= absent). For a given itemset s, the technique identifies among the rules with antecedents subsumed by s those that have the highest precedence according to the reliability of the rules—this liability is assessed based on the rules' confidence and support values.

The rule is then used for the prediction of ij . The method suffers from three shortcomings. First, it is clearly not suitable in domains with many distinct items ij. Second, the consequent is predicted based on the"testimony" of a single rule, ignoring the simple fact rules with the same antecedent can imply different consequents—a method to combine these rules is needed. Third, the system may be

sensitive to the subjective user- specified support and confidence thresholds Some of these weaknesses are alleviated in [11], where a missing item is predicted in four steps. First, they use a so-called partitioned-ARM to generate a set of association rules (a ruleset). The next step prunes the ruleset (e.g., by removing redundant rules). From these, rules with the smallest distance from the observed incomplete shopping cart are selected. Finally, the items predicted by these rules are weighed by the rules' antecedents' similarity to the shopping cart.

The approach in [10] pursues a Dempster-Shafer (DS) belief theoretic approach that accommodates general data imperfections. To reduce the computational burden, Hewa- wasam et al. employ a data structure called a belief itemset tree. Here, too, rule generation is followed by a pruning algorithm that removes redundant rules. In order to predict the missing item, the technique selects a "matching" ruleset—a rule is included in the matching rule set if the incoming itemset is contained in rule antecedent.

If no rules satisfy this condition, then, from those rules that have nonempty intersection with the item set s, rules whose antecedents are "closer" to s according to a given distance criterion (and a user-defined distance threshold) are picked. Confidence of the rule, its "entropy," and the length of its antecedent are used to assign DS theoretic parameters to the rule. Finally, the evidence contained in each rule belonging to the matching rule set is combined or "pooled" via a DS theoretic fusion technique.

In principle, at least, we could adopt any of the above methodologies; but the trouble is that they were all designed primarily for the classification task and not for shopping cart completion. Specifically, the number of times such classifiers have to be invoked would be equal to the number of all distinct items in the database (i.e., n) minus the number of those already present in the shopping cart. This is why we sought to develop a predictor that would predict all items in a computationally tractable manner.

Another aspect of these approaches is the enormous amount of effort/cost it takes to obtain a tangible and meaningful set of rules. The root of the problem lies in the a priori-like algorithms used to generate frequent item sets and the corresponding association rules—the costs become prohibitive when the database is large and complicated.

Here, the size and difficulty are determined by four parameters: number of transactions, number of distinct items, average transaction length, and the minimum support threshold. For example, the problem can become intractable if the number of frequent items is large; and whether an item is frequent or not is affected by the minimum support threshold. It is well known that a priori-based algorithms suffer from performance degradation in large-scale problems due to combinatorial explosion and repeated passes through the database [11].

4. RESEARCH AND FINDINGS

a) Data Mining

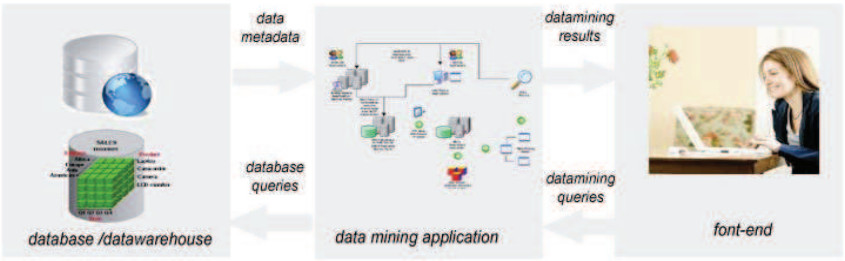

Figure 1:- Data Mining Architecture

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.[1] These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

b) Data Mining Techniques

Five Techniques in Data Mining
Association
Classification
Clustering
Prediction
Sequential

c) Data mining Architecture

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses and etc[10]. This question leads to four possible architectures of a data mining system as follows:

No Coupling
Loose Coupling
Semi Tight Coupling
Tight Coupling

5. PROPOSED SYSTEM ANALYSIS AND DESIGN

a)Analysis

Proposed system:

- ✓ Finding missing symptoms using apriority algorithms in frequently used symptoms list.

- ✓ Counting number of users per symptoms.
- ✓ Calculating total number of visitor's in our websites

Advantages of Proposed system:

- ✓ Reducing time complexity. User can easily view the symptoms.
- ✓ Missing symptoms can easily find in the symptoms list.

B) Problem Statement

a) Let I={i1,…….in} be a set of distinct symptoms.

b) Let a database consist of transactions T1,…Tn such that Ti c I, $\forall$i

c) Let X be a group of symptoms such that X c I

d) An association rule has the form r(a) => r(c)

e) where , r(a) and r(C) are the set of symptoms.

f) The r(a) is the antecedent and r(c) is consequent.

g) The rule reads:- If all symptoms from r(a) are present in a transaction from r(c) are also present in same transaction.

h) This rule does not have absolute reliability.

i) The probabilistic confidence in rule r(a) => r(c) Can be defined with the help of supports (relative frequencies) of antecedents and consequent as the percentage o transaction that contain r(c) among those transactions that contain r(a) :

j) Conf = support (r(a) U r(c) / support r(a) ----------(1)

k) Let us assume that an association mining program has already discovered all high support set of symptoms.

l) For each such symptom ,X, any pair of subsets r(a) and r(c) such that

m) r(a) U r(c) = X and r(a) $\cap$ r(c) =Ø

n) We can define an association rule

o) r: r(a) => r(c)

p) The number of rules implied by X grows exponentially in the number of symptoms in X

q) We usually consider only high confidence rules derived from high support list of symptoms [4].

C) Detail designed

Data Flow Diagrams:

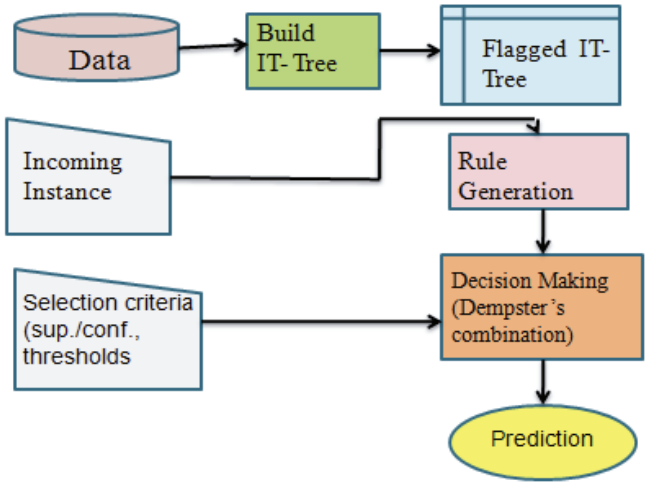

Fig.2. Data flow of our proposed system

D) The Bayesian approach

- In this approach suppose we want to establish the presence of list of symptoms, [5],[6], $S=\{i_1^{(s)},\ldots,i_k^{(s)}\}$ increases the chance that item $i_j \in S$ is also present.

- Bayes rule yields

$$P(i_j \mid i_1^{(s)},\ldots,i_k^{(s)}) = P(i_1^{(s)},\ldots,i_k^{(s)} \mid i_j)\, P(i_j)$$

$$P(i_1^{(s)},\ldots,i_k^{(s)})$$

- We select all the symptoms for which; $P(i_j \mid i_1^{(s)},\ldots,i_k^{(s)}) > P(\neg i_j \mid i_1^{(s)},\ldots,i_k^{(s)})$

  where, $P(\neg i_j \mid i_1^{(s)},\ldots,i_k^{(s)})$ is the probability of the symptom i_j being absent given the list of symptoms S is present.

- Since, the denominator is the same for any given list of symptoms S, it is enough if the classifier chooses the symptoms that maximizes the value of numerator.

E) The Proposed Solution

The proposed Rule Generation Algorithm makes use of the flagged IT Tree created from the training data set.

- The algorithm takes an incoming list of symptoms as the input and returns a graph that denes association rules entitled by the incoming list of symptoms [3],[4].

Itemset Tree Construction (IT Tree)

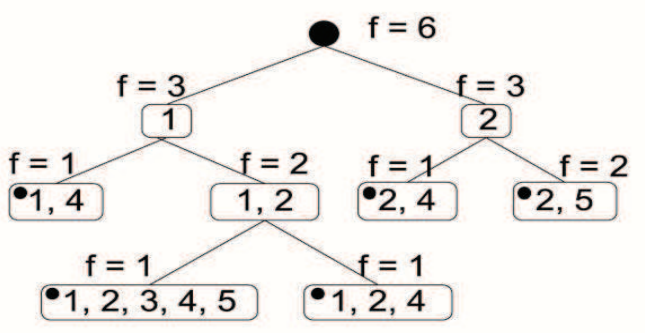

Fig. 3. The IT-tree constructed from the database .

F) Employing the DS Theory

When searching for a new way to predict the presence or absence of an item i_j in a partially observed medical disease S, we wanted to use association rules. However many rules with equal antecedents differ in their consequents- some of these consequents contain i_j , others do not. The question is how o combine the potentially conflicting evidence. One possibility is to rely on the DS theory of evidence combination. Let us now describe our technique, which we refer to by the acronym DS-ARM (Dempster-Shafer-based Association Rule Mining).

6. REFERENCES

1. S. Noel, V.V. Raghavan, and C.H. Chu, *"Visualizing Association Mining Results through Hierarchical Clusters,"* Proc. Int'l Conf. Data Mining (ICDM '01) pp. 425-432, Nov./Dec. 2001.
2. P. Bollmann-Sdorra, A. Hafez, and V.V. Raghavan, *"A Theoretical Framework for Association Mining Based on the Boolean Retrieval Model,"* Data Warehousing and Knowledge Discovery: Proc. Third Int'l Conf. (DaWaK '01), pp. 21-30, Sept. 2001.
3. R. Agrawal, T. Imielinski, and A. Swami, *"Mining Association Rules between Sets of Items in Large Databases,"* Proc. ACM Special Interest Group on Management of Data (ACM SIGMOD), pp. 207-216, 1993.
4. M. Kubat, A. Hafez, V.V. Raghavan, J.R. Lekkala, and W.K. Chen, *"Itemset Trees for Targeted Association Querying,"* IEEE Trans. Knowledge and Data Eng., vol. 15, no. 6, pp. 1522-1534, Nov./Dec. 2003.
5. A. Rozsypal and M. Kubat, *"Association Mining in Time-Varying Domains,"* Intelligent Data Analysis, vol. 9, pp. 273-288, 2005.
6. V. Raghavan and A. Hafez, *"Dynamic Data Mining,"* Proc. 13th Int'l Conf. Industrial and Eng. Applications of Artificial Intelligence and Expert Systems IEA/AIE, pp. 220-229, June 2000.

10. C.C. Aggarwal, C. Procopius, and P.S. Yu, *"Finding Localized Associations in Market Basket Data,"* IEEE Trans. Knowledge and Data Eng., vol. 14, no. 1, pp. 51-62, Jan./Feb. 2002.

11. R. Bayardo and R. Agrawal, *"Mining the Most Interesting Rules,"* Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 145-154, 1999.

12. J. Zhang, S.P. Subasingha, K. Premaratne, M.-L. Shyu, M. Kubat, and K.K.R.G.K. Hewawasam, *"A Novel Belief Theoretic Association Rule Mining Based Classifier for Handling Class Label Ambiguities,"* Proc. Workshop Foundations of Data Mining (FDM '04), Int'l Conf. Data Mining (ICDM '04), Nov. 2004.

****

---

[1]*Author 1: Mrs. Shweta A.Gode, Lecturer Dept (CSE),DMIETR, Sawangi (M), Wardha (M.S)*
shweta_amt80@rediffmail.com
[2]*Author 2: Dr. G.R.Bamnote Professor, CSE, PRMITR, Badnera, Amravati (M.S)*
grbamnote@rediffl.com