2012 International Conference on Information and Computer Networks (ICICN 2012)
IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore

Mining Association Rules from Large Datasets Towards Disease Prediction

K.Srinivas [1], G.Raghavendra Rao [2] and A.Govardhan [3 +]

[1] Associate Professor, Jyothishmathi Institute of Technology& Science, Karimnagar, India
[2]Principal, NIE Institute of Technology & Science, Mysore
[3] Professor & Principal JNTUH College of Engineering, Karimnagar, India

Abstract. In data mining association rule mining represents a promising technique to find hidden patterns in large data bases. The main issue about mining association rules in a medical data is the large number of rules that are discovered, most of which are irrelevant. Such number of rules makes the search slow. However, not all of the generated rules are interesting, and some rules may be ignored. In medical terms, association rules relate disease data measures the patient risk factors and occurrence of the disease. Association rule medical significance is evaluated with the usual support and confidence metrics. Association rules are compared to predictive rules mined with decision trees, a well-known machine learning technique. In this paper we propose a new and simple measure to find the strength of association among the attributes of a given attribute set and find the occurrence of the disease based on the attribute association.

Keywords: Data mining; association strength; minimum support; minimum confidence;

1. Introduction

With the ever-growing complexity in recent years, huge amounts of information in the area of medicine have been saved every day in different electronic forms such as Electronic Health Records (EHRs) and registers. These data are collected and used for different purposes. Data stored in registers are used mainly for monitoring and analyzing health and social conditions in the population. The unique personal identification number of every inhabitant enables linkage of exposure and outcome data spanning several decades and obtained from different sources. The existence of accurate epidemiological registers a basic prerequisite for monitoring and analyzing health and social conditions in the population. Some registers are state-wide, cover the whole collieries population, and have been collecting data for decades. They are frequently used for research, evaluation, planning and other purposes by a variety of users in terms of analyzing and predicting the health status of individuals.

1.1. Focus of the survey

Recent studies have documented poor population health outcomes in coal mining areas. These findings include higher chronic cardiovascular disease (CVD) mortality rates and higher rates of self-reported CVD [13]. The risk for CVD is influenced by environmental, genetic, demographic, and health services variables. Risk behaviors, in turn, are related to lower socio economic status (SES); low SES persons are more likely to smoke, consume poor quality diets, and engage in sedentary lifestyles. Coal mining areas are characterized by lower SES relative to non-mining areas, suggestive of higher CVD risk. Environmental agents that contribute to CVD include arsenic, cadmium and other metals, non-specific particulate matter (PM), and polycyclic aromatic hydrocarbons (PAHs). All of these agents are present in coal or introduced into local

+ Corresponding author. Tel.: +918782236414; fax: +918782245821.
E-mail address: jaya_konda@yahoo.com

ambient environments via activities of coal extraction and processing. Most previous research on population health in coal mining areas has employed state-level mortality data rather than individual-level data. An exception was a study of self-reported chronic illness in relation to coal mining; this study was limited to a non-standard assessment instrument with limited individual-level covariates in Singareni Collieries, Andhra Pradesh state in country India. The current study uses Area Hospitals data to assess CVD risk in coal mining areas before and after control for individual-level covariates including smoking, obesity, co-morbid diabetes, alcohol consumption and others. We propose to test the association among the co-morbid attributes as which attribute supports that CVD rates will be significantly elevated for residents of coal mining regions after controlling for covariates, suggestive of an environmental impact.

1.2. Data Mining concepts in Health Care

Data Mining aims at discovering knowledge out of data and presenting it in a form that is easily compressible to humans. It is a process that is developed to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. In practice, the two primary goals of data mining tend to be *prediction* and *description* [14][1]. *Prediction* involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description,* on the other hand, focuses on finding patterns describing the data that can be interpreted by humans.

2. Basic concepts and terminology

This section introduces association rules terminology and some related work on rare association rules.

2.1 Association Rules

Formally, association rules are defined as follows: Let I = {i_1, i_2, …., in} be a set of items, *D* be a set of transactions, where each transaction *T* is a set of items such that $T \subseteq I$. Each transaction is associated with a unique identifier *TID*. 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* $X \cap Y = \Phi$. The rule $X \rightarrow Y$ has *support s* in the transaction set *D* if *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 than given thresholds. These rules are called *strong rules*, and the framework is known as the support-confidence framework for association rule mining.

2.2 Transforming Medical Data Set

A medical dataset with numeric and categorical attributes must be transformed to binary dimensions, in order to use association rules. Numeric attributes are binned into intervals and each interval is mapped to an item. Categorical attributes are transformed by mapping each categorical value to one item. Our first constraint is the negation of an attribute, which makes search more exhaustive. If an attribute has negation then additional items are created corresponding to each negated categorical value or each negated interval.

Missing values are assigned to additional items, but they are not used. In short, each transaction is a set of items and each item corresponds to the presence or absence of one categorical value or one numeric interval.

3. Proposed Work

We propose the AA(I) Attribute Association which is an extension to OA[3] which finds the association among the attributes[4] of a dataset. A patient having disease that can be always a combination of symptoms

such as fever may come with stress or due to change in climate. The other patient may have fever with cold and cough. Our interest is to find the strength between the symptoms or diseases how frequently they are associated. In our future study we would like to extend this to the heart attack and find the strength between co-morbid attributes influencing the patient towards CVD.

Let I = $\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_m\}$ be an attribute set. The association of attribute can be denoted defined as follows:

$$AA(I) = \sum_{I' \subseteq I, |I'| \geq 2} \frac{s(I' \neg I'')}{Total\ no\ of\ transactions} \frac{|I'|}{|I|}$$

Where $I' \subseteq I$ and $I'' = I - I'$

$$AA(I) = \begin{cases} 0 - no\ association \\ < \propto - weak\ association \\ \geq \propto - strong\ association \end{cases}$$

ALGORITHM: DAST (Defining Association Strength)

Input: TD-transaction Database, MS-Minimum Support, MC-Minimum Confidence, MA-Minimum Association

Output: Strong Association datasets

Method:

1. C_1= Candidate 1-itemsets
2. L_1=frequent 1-itemsets
3. for(K=2;$L_{K-1} \neq \varphi$; K++)
4. {
5. $C_K = L_{K-1} \bowtie L_{K-1}$
6. for each c ε C_K
7. If any subset of c ∉ L_{K-1}
8. then $C_K = C_K - \{c\}$
9. for each c ε C_K
10. { If support(c) ≥ Ms then $L_k = L_K$ U {c} }
11. for each c ε L_k
12. {
13. If A(c)≥ Ma
14. {
15. for each c=(x U y) *// x contains any number of items but y contains only one item //*
16. If confidence(x → y) ≥ Mc
17. then SAR=SAR U { x→y}
18. }
19. }
20. }

4. Experimental Results

We have conducted experiments by using our algorithm DAST on two data sets by taking the *support* as 20% and by taking the *threshold value* for α as 25.

The first data is a synthetic dataset.

{{ f,d,b} {a,b,c,f} {b,e,f} {a,f,c} {f,b,c} {c,a} {d} {b} {a,c,b} { c,e,a }}

By applying our algorithm DAST we get

C_3		L_3
abc	2	abc
abf	1	acf
acf	2	bcf
bcf	2	

We have calculated the associations among the attributes by using our measure Association Attributes AA for every stage of candidates i.e. C_2, C_3.

The measure uses negative association rule and find the association among the items with the absence of one a item in the given set I and finds the association among the items in the itemset I.

$|I|$ denotes *the total number of attributes*
$|I'|$ *denotes the subset of I should be* ≥ 2
T *denotes the total transactions in the dataset*
$\neg \lambda$ *denotes the absence of attribute* λ *in the given dataset I*

The measure can be expanded as below for all the itemsets in C_3 and finds the AA (I) as below for instance I = abc then our measure can be defined as

$$AA(I) = \frac{S(ab\neg \text{c})}{T} X \frac{I'}{I} + \frac{S(bc \neg \text{a})}{T} X \frac{I'}{I} + \frac{S(ac\neg \text{b})}{T} X \frac{I'}{I} + \frac{S(\text{abc})}{T} X \frac{I'}{I}$$

We get the values for

AA (abc) = **04**
AA (acf) = **0.47**
AA (bcf) = **0.47**

Where the values are greater than threshold value α which is given as 0.25 and according to the associations among the attributes the attributes has moderate support among each other.

The same will be converted among the dataset of diseases where a person is suffering from cold, fever and other related symptoms. The real time data set of seasonal fever is collected from the local doctors of Singareni Collieries which consist of six attributes as {cold, headache, fever, bodypain, allergy, cough}. Different patients may have the different combination of symptoms.

The following example data set in Table1 show the data of diseases.

TABLE I. DISEASE DATASET

01	cold, fever, allergy
02	cold, headache, cough
03	cold, headache, bodypain, fever
04	fever, bodypain, cough
05	cold, fever, headache, cough
06	cold, bodypain, cough
07	cold, allergy, cough
08	cold ,cough, bodypain
09	cold, cough
10	cold, headache, fever

We applied our algorithm to find the association among the attributes with dataset having approximately 1000 records. Finally the most frequent data items among the total transactions are calculated. The threshold value α is fixed as 0.25.

AA (cold, headache, fever) get the association among the attributes as 0.26.

AA (cold, bodypain, cough) get the association among the attributes as 0.53

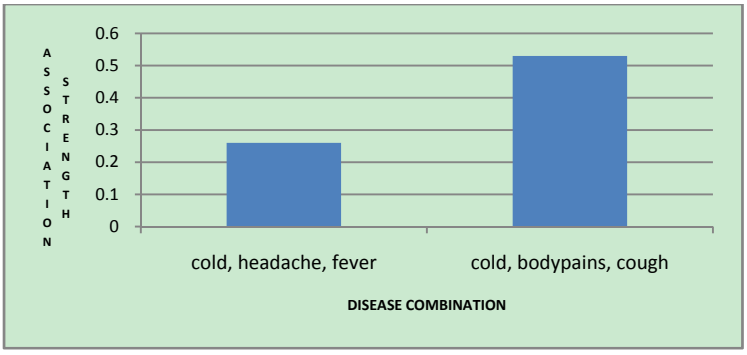

Fig 1. Comparison of association of attributes

The dataset (cold, headache, fever) gives association as 0.26 which is very close to the threshold value which is 0.25. We can say α ≥ 0.26.

The combination in the dataset with (cold, body pains, and cough) gives the association among the attributes as 0.53 which is much greater than the threshold value as show in fig1. We can say this dataset contains strong association among these attributes.

In the terms of medical we can clearly say that {cold, body pains, cough} are closely associated to each other, which indicates a person suffering from cold will have body pains and cough

Here we applied our algorithm to small and real dataset which proved to be satisfactory. Further this need to be extended to the larger databases [10] and other diseases like heart attack.

5. Conclusion and Future Work

In this paper, we proposed an algorithm that mines the association among the various attributes in a dataset. Our method generates valid association rules by taking a probability measure. We conducted experiments on synthetic and real data sets. We have applied the measure to both frequent and infrequent itemset to the dataset1. Surprisingly we found that the infrequent itemset is also having the association among the attributes. This type of association is possible in the case of diseases. A patient may have some disease and that can be treated with rare symptoms like 18 year old young boy is getting heart attack. In our future work we wish to conduct experiments on large real time health datasets to predict the diseases like heart attack and compare the performance of our algorithm with other related algorithms.

6. References

[1] Mannila, H.: Methods and Problems in Data Mining. In: The International Conference on Database Theory, pp. 41–55 (1997)

[2] Jiawei, H., Jian, P., Yiwen, Y., Runying, M.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 53–87 (2004)

[3] Animesh Adhikari, P.R. Rao, Capturing association among items in a database, Data & Knowledge Engineering 67 (2008) 430–443

[4] Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In:ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, pp. 337–341 (1999)

[5] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, DC, 1993, pp. 207–216.

[6] Szathmary, L., Napoli, A., Valtchev, P. Towards rare itmeset mining. In *International Conference on Tools with Artificial Intelligence*, Washington, DC. 2007, pp. 305-312.

[7] Chia-Wen Liao , Yeng-Horng Perng, Tsung-Lung Chiang Discovery of unapparent association rules based on extracted probability, Journal Decision Support Systems Volume 47 Issue 4, November, 2009

[8] Yun Sing Koh1 Russel Pears Rare Association Rule Mining via Transaction Clustering Seventh Australasian Data Mining Conference (AusDM 2008), Glenelg, Australia.

[9] S.L. Hershberger, D.G. Fisher, Measures of Association (Encyclopedia of Statistics in Behavioral Science), John Wiley & Sons, 2005

[10] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of SIGMOD Conference on Management of Data, 1993, pp. 207–216.

[11] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules, in: Proceedings of Knowledge Discovery in Databases, 1991, pp. 229–248

[12] Hendryx and Ahern, 2008, Chronic Illness Linked To Coal-Mining Pollution, Study, ScienceDaily , 2008

[13] Efficient Discovery of Risk Patterns in Medical Data,case study Jiuyong Li, Ada Wai-chee Fu, Paul Fahey, Artificial Intelligence in Medicine (2008)

[14] Evaluating association rules and decision trees to predict multiple target attributes, Carlos Ordonez and Kai Zhao, Intelligent data Analysis 15 (2011) 173–192 173,DOI 10.3233/IDA20100462,IOS Press