Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 8, August 2013, pg.91 – 95

RESEARCH ARTICLE

Analysis of Attribute Association Rule from Large Medical Datasets towards Heart Disease Prediction

D. P. Shukla[1], Shamsher Bahadur Patel[2], Ashish Kumar Sen[3], Pramod Kumar Yadav[4]

[1]Department of Computer Science & Mathematics, Govt. P.G. Science College Rewa (M.P.), India
[2]Department of Computer Science & Mathematics, Govt. P.G. Science College Rewa (M.P.), India
[3]Department of Computer Science & Mathematics, Govt. P.G. Science College Rewa (M.P.), India
[4]Department of BCA & Physics, Govt. P.G. Science College Rewa (M.P.), India

[1]shukladpmp@gmail.com; [2]sstpatel12@gmail.com; [3]ashishsen1983@gmail.com; [4]yadav.pramod181@mail.com

Abstract— Cardio vascular disease is a major threat to half of the world population. The term heart disease is related to all the diverse diseases affecting the heart. The healthcare industry generates large amount of data that are too difficult to be analyzed by traditional methods. Hence computer assisted methods are necessary to make correct decisions. Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. 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. A rule-based decision support system (DSS) is presented for the diagnosis of coronary vascular disease (CVD). 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 rules are compared to predictive rules mined with decision trees, a well-known machine learning technique. In this paper we propose a new system to find the strength of association among the attributes of a given data set. The proposed system has several advantages since it is automatically generated. It provides CVD diagnosis based on easily and none invasively acquired features.

Key Terms: - Data mining; Association Rule Mining; Decision Tree; Machine Learning Technique

I. INTRODUCTION

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Healthcare industry today generates large amount of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making.

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).

We also refer to Computer-aided diagnosis methodologies as stated [15] in the literature; in this case, the data obtained by some of the aforesaid methods or other sources (i.e., laboratory examinations, demographic and/or history data, etc.) are evaluated from a computer-based application, leading to a CVD diagnosis. These methodologies can be divided into various categories, based on the type of data they use for subject characterization: 1)methods that employ the resting or exercise ECG of the patient, extracting features from it, such as the ST segment[17], [16],, the QT interval , the T wave amplitude , the R wave, and the heart rate variability (HRV) ; 2) methods using medical images such as SPECT ; 3) methods based on heart sounds associated with coronary occlusions[18] ; 4) methods based on arterio-scillography[19]; 5) methods based on Doppler ultrasound signals[20] ; 6) methods employing demographic, history, and laboratory data (subject's data) ; and 7) methods combining more than one type of data such as ECG, scintigraphy, and subject's data.

II. BASIC CONCEPTS AND TERMINOLOGY

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

A. *Association Rules:*

Formally, association rules are defined as follows: Let I = {i1, *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* contains *X U 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.

B. *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.

Markos G. Tsipouras et.al proposed as in prediction of heart attack describes two demographic features were recorded: the age and sex of the patient. From the subject's history, the family history of CAD (FH), smoking history (Smok), history of diabetes mellitus (DM), and hypertension (HT) or hyperlipidaemia were used. Family history of CAD was defined as the presence of CAD in the father or brother aged <55 years or mother or sister aged <65 years. Current and ex-smokers were defined as having smoked the last cigarette less than a week and less than a year before CA, respectively. Diabetes mellitus was defined as a fasting blood glucose concentration (FBGC) ≥126 mg/dl or antihyperglycemic drug treatment, hypertension as systolic blood pressure (SBP) >140 mmHg, and/or diastolic blood pressure (DBP) >90 mmHg or use of antihypertensive agents, and hyperlipidemia as fasting total cholesterol >220 mg/dl or use of lipid-lowering agents (statins or fibrates). Other clinical data were also recorded; body mass index (BMI), calculated as weight (kg) divided by the square of height (square meter), waist perimeter measured in centimeter, resting heart rate (HR), measured in beats per minute (b/min), resting SBP and DBP measured in mmHg. The laboratory investigations also incorporated were creatinine (Cre), glucose (Glu), total cholesterol (Tchol), high-density lipoprotein (HDL), and triglycerides (TRG) measured in milligrams per deciliter (mg/dL). All the aforementioned features are considered to be traditional cardiovascular risk factors widely used to assess the risk of CAD. In addition, carotid–femoral pulse wave velocity (PWVcf)

and augmentation index (AIx) expressed in meter per second and percentage, respectively, were also used as noninvasive indices of arterial stiffness.

In the association rules we generate the rules such as $Rule_i$= (at_1 op V1) ∧ (at_2 op V2) ∧ · · ·∧ (at_m op Vm), where (at, Vj) is a attributes and its threshold-values pair and op is a comparison operator can be (=, !=, >,<,, ≤, ≥).

III. 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 = {λ1, λ2, λ3,. . ., λm }be an attribute set. The association of attribute can be denoted defined as follows:

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

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

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

Further we calculate frequencies of various attributes in the dataset which has more association among the attributes and analyze the results.

ALGORITHM: DAST (Defining Association Strength)

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

Output: **Strong Association datasets**

Method:

Step 1. C1= Candidate 1-itemsets
Step 2. L1=frequent 1-itemsets
Step 3. for (K=2; LK-1 ≠ φ; K++)
Step 4. {
Step 5. CK= LK-1 ⋈ LK-1
Step 6. for each c ε CK
Step 7. If any subset of c ∉ LK-1
Step 8. Then CK = CK - {c}
Step 9. for each c ε CK
Step 10. {
Step 11. If support(c) ≥ Ms then Lk= LK U {c}
Step 13. }
Step 14. for each c ε Lk
Step 15. {
Step 16. If A(c) ≥ Ma
Step 17. {
Step 18. for each c=(x U y) *// x contains any number of items but y contains only one item //*
Step 19. If confidence(x → y) ≥ Mc
Step 20. then SAR=SAR U { x→y}
Step 21. }
Step 22. }
Step 23. }

IV. EXPERIMENT AND RESULT

We calculate frequencies of various attributes in the dataset which has more association among the attributes and analyze the results.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	25-35	7	2.3	2.3	2.3
	35-45	57	18.6	18.6	20.9
	45-55	89	29.0	29.0	49.9
	55-65	121	39.4	39.4	89.3
	65-75	31	10.0	10.0	99.3
	>75	2	0.7	0.7	100.0
	Total	307	100.0	100.0	

Table A. For Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	98	32.0	32.0	32.0
	Female	209	68.0	68.0	100.0
	Total	307	100.0	100.0	

Table B. For Sex

The above tables show that the CVD risk is more in male gender that are in the age range between 55 & 65. Similar measures are applied to calculate the statistics of each attributes frequency and we can apply this for prediction of heart attack.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new measure that finds 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 dataset. 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. 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.

REFERENCES

[1] Mai Shouman, Tim Turner, Rob Stocker,(2012),"Using Data Mining Techniques In Heart Disease Diagnosis And Treatment ",Proceedings in Japan-Egypt Conference on Electronics, Communications and Computers,IEEE,Vol.2 pp.174-177.

[2] K.Srinivas , B.Kavita Rani, Dr. A.Govardhan (2010), Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks ,IJCSE Vol. 02, No. 02, pp 250-255.

[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 itemset 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 26
[15] Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis
[16] J. W. Deckers, B. J. Rensing, R. V. H. Vinke, and M. L. Simoons, "Comparison of exercise algorithms for diagnosis of coronary artery disease,"in Proc. Comput. Cardiology, 1988, pp. 231–234.
[17] K. Lewenstein, "Radial basis function neural network approach for the diagnosis of coronary artery disease based on the standard electrocardiogram exercise test," Med. Biol. Eng. Comput., vol. 39, pp. 1–6, 2001.
[18] Y. M. Akay, M. Akay, W. Welkowitz, J. L. Semmlow, and J. Kostis, "Noninvasive acoustical detection of coronary artery disease: A Comparative study of signal processing methods," IEEE Trans. Biomed. Eng., vol. 40, no. 6, pp. 571–578, Jun. 1993
[19] M. Pouladian,M. R. H. Golpayegani, A. A. Tehrani-Fard, andM. Bubvay-Nejad, "Noninvasive detection of coronary artery disease by arteriooscillography,"IEEE Trans. Biomed. Eng., vol. 52, no. 4, pp. 743–747,Apr. 2005. [20] I. Guler and E. D. U beyli, "Automated diagnostic systems with diverse and composite features for Doppler ultrasound signals," IEEE Trans. Biomed. Eng., vol. 53, no. 10, pp.1934–1942, Oct. 2006.

