K.Srinivas, G.Raghavendra Rao, A.Govardhan / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue4, July-August 2012, pp.1680-1683

Analysis of Attribute Association in Heart Disease Using Data Mining Techniques

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 in CSE, JNTUH, Hyderabad, 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. A rule-based decision support system (DSS) is presented for the diagnosis of coronary vascular disease (CVD). The dataset used for the DSS generation and evaluation consists of 1897 subjects, each one characterized by 21 features, including demographic and history data, as well as laboratory examinations. 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 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 noninvasively acquired features, and also able to provide interpretation for the decisions made.

I. 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 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.

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.

A. 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.

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 = {i_1, i_2, …., i_n} 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.

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.

Carlos Ordonez et al proposed the data as specified. Transformation parameters default values we must discuss attributes corresponding to heart vessels. The LAD, RCA, LCX and LM numbers represent the percentage of vessel narrowing (stenosis) compared to a healthy artery, where narrowing is 0%. Attributes LAD, LCX and RCA were binned at 50% and 70%. In cardiology a 70% value or higher indicates significant stenosis and a 50% value indicates borderline disease. Stenosis below 50% generally indicates the patient is considered healthy since it is unlikely the artery may get blocked. The LMartery has a lower (more stringent) cutoff because it poses a higher health risk than the other three arteries. The fundamental reason

is LAD and LCX arteries branch from LM. Therefore, a defect in LM is likely to trigger more severe disease. Attribute LM was binned at 30% and 50%. The 9 heart regions (AL, IL, IS, AS, SI, SA, LI, LA, AP) were partitioned into 2 ranges at a cutoff point of 0.2, meaning a perfusion measurement greater or equal than 0.2 indicated a severe defect. CHOL was binned at 200 (warning) and 250 (high). AGE was binned at 40 (adult) and 60 (old). Finally, only the four artery attributes (LAD, RCA, LCX, LM) had negation to find rules referring to healthy patients and sick patients. The other attributes did not have negation. The data set had nulls values. We applied a common missing value imputation solution: for numeric attributes we used the mean and missing categorical values were substituted by the mode of the attribute.

In the association rules we generate the rules such as $Rule_i = (at_1 \text{ op } V_1) \wedge (at_2 \text{ op } V_2) \wedge \cdots \wedge (at_m \text{ op } V_m)$, where (at, V_j) is a attributes and its threshold-values pair and op is a comparison operator can be (=, !=, >,<,, ≤, ≥).

Some common attributes are verified and the results are observed such as age > 60 and bp >180 and chol > 50 then the patient is in risk zone and we can also predict heart attack. Such common attributes can be associated and a measure is generated to find how frequent the association is occurring among the attributes. When the large data sets are processed an alternative method can be used to reduce the time to analyze the dataset by choosing the common high associated attributes and the results can be considered.

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 = $\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_m\}$ 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 $I'' = I - I'$

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

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

age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	25-35	7	2.3	2.3	2.3
	36-45	56	18.5	18.5	20.8
	46-55	88	29.0	29.0	49.8
	56-65	119	39.3	39.3	89.1
	66-75	31	10.2	10.2	99.3
	>75	2	.7	.7	100.0
	Total	303	100.0	100.0	

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	97	32.0	32.0	32.0
	Male	206	68.0	68.0	100.0
	Total	303	100.0	100.0	

The above tables show that the CVD risk is more in male gender whoare in the agre range between 56 & 65. Similar measures are applied to calculate the statistics of each attribtes frequency and we can apply this for prediction of heart attack.

IV. 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 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.

Reference

[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 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.