International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 17 Issue III, December 2019

COMBINED METHODOLOGY of the CLASSIFICATION RULES for MEDICAL DATA-SETS

Dhanush K, Pavitra Reddy

Abstract-'Data mining' offers methodological and technical solutions to deal with the analysis of medical data and construction of prediction models. A large variety of these methods requires general and simple guidelines that may help practitioners in making clinical decisions. The purpose of this study was to build a hybrid data mining model to extract classification knowledge for various health hazards to aid in clinical decisions in an emergency department. This study utilized real world data collected from an emergency department of a hospital and used a new model which is developed combining the Apriori algorithm and a C5.0 algorithm to generate a classification rule base for the classification of medical data-sets, which can help physicians to make clinical decisions faster and more accurately.

Keywords: Data mining, chest pain, clinical decisions, emergency department, Hybrid model

1. INTRODUCTION

Data mining and knowledge discovery in databases have been attracting significant amount of research in fields like industry, medical, commerce, science which is attracting the media attention of late .In this paper we are going to concentrate on data mining in medical field problems have been caused from several internal and external factors, including patient characteristics, staffing patterns of emergency department, access to health care providers, patient arrival time, management practices, and testing and treatment strategies selected by emergency Department. Emergency departments serve many functions in the current healthcare system, including initial management of patients with critical illnesses and primary care for a growing proportion of the population. Overcrowding of emergency departments is a growing problem. Delays in admitting patients to inpatient units have been reported as a contributing factor to overcrowding. This study to examine the incidence of critical illness in the emergency department and its total burden as reflected in emergency department length of stay. The goal of predictive data mining in clinical medicine is to derive models that use patient information to support specific clinical decisions. Data mining models can be applied to building of decision-making procedures such as prognosis, diagnosis, and treatment planning, which once evaluated and verified, could then be embedded in clinical information systems.

2. METHODOLOGY REVIEW

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst. Medical data mining has been applied to accurate classification and rapid prediction for prognosis and diagnosis of patients in a specialized medical area. It has been also used for training unspecialized doctors to solve a specific diagnostic problem. Data mining tasks can, in general, be classified to tasks of description and prediction. Predictive data mining in clinical medicine deals with learning models to predict patients' health. The models can be devoted to support clinicians in diagnostic, therapeutic, or monitoring the tasks. Yun (2008) utilized a C4.5 algorithm to build a decision tree in order to discover the critical causes of type II diabetes. She has learned

about the illness regularity from diabetes data, and has generated a set of rules for diabetes diagnosis and prediction. Khan et al. (2009) used decision trees to extract clinical reasoning in the form of medical

International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 17 Issue III, December 2019

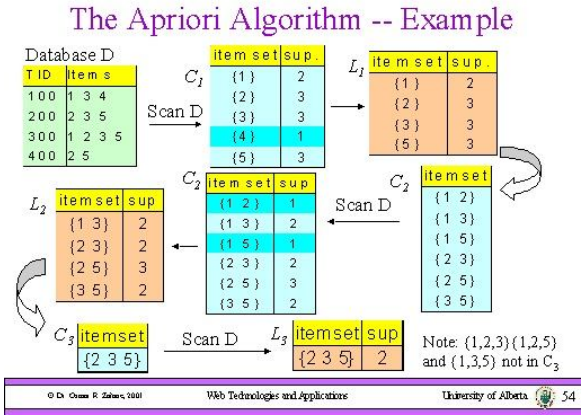
expert's actions that are inherent in a large number of electronic medical records. The extracted data could be used to teach students of oral medicine a number of orderly processes for dealing with patients with different problems depending on time. Tan et al. (2007) used the Apriori algorithm to mine the rules for the compatibility of drugs from prescriptions to cure arrhythmia in the traditional Chinese medicine database. The experimental results showed that the drug compatibility obtained by the Apriori algorithm is generally consistent with the traditional Chinese medicine for that disease. Abdullah et al. (2008) adopted an association algorithm to find the relationship between diagnosis and prescription. They stated that purchases and medical bills have much in common. Therefore, the Apriori algorithm was useful to figure out large item sets and to generate association rules in medical billing data.

3 CLASSIFICATION RULES

The classification rules in the medical data mining include the application of the various classifiers on the data-sets. In our study we are going to develop a hybrid methodology in which Aproiri algorithm is used on the data item which includes the various lab test that the patient has to undergo in order to give the appropriate diagnosis. Then the C5.0 algorithm is used on the data-sets.

3.1 APRIORI ALGORITHM

Developed by Agarwal and Srikant in 1994.Innovative rule to find association rules on large scale allowing implication outcomes that consist of more than one items. Apriori is designed to operate on databases containing transactions (for example, collections of lab tests made by patients, or details of a patient's medical web history frequentation) is common in association rule mining, given a set of *item sets* (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found



On minimum support value. Consider the same above example of a database D, consisting of 4 transactions. Suppose min. support count required is 2 (i.e. min_sup = 2/4 = 50 %) The first scan of database is same as Apriori, which derives the set of 1-itemsets & their support counts. The set of frequent items is sorted in the order of descending support count. The resulting set is denoted as L = {I2:3, I3:3, I5:3, I1:2} Usefulness of a rule can be measured with a minimum SUPPORT threshold. Database D consists of events T_1, T_2, T3,that is D = {T_1, T_2, T_3}Let there be an itemset X that is a subregion of event T_k, that is $X \subseteq T_k$. The support can be defined as

$$SUP(X) = \frac{|\{T_k \in D \mid X \subseteq T_k\}|}{|D|}$$

This relation compares number of events containing itemset X to number of all events in databaseCertainty of a rule can be measured with a threshold for confidence.This parameter lets to measure how often an event's itemset that matches the left side of the implication in the association rule also matches for the right side .Rules for events whose itemsets do not match sufficiently often the right side while mathching the left (defined by a threshold value) can be excluded. If confidence gets a value of 100 % the rule is an exact rule. Even if confidence reaches high values the rule is not useful

unless the support value is high as well.Rules that have both high confidence and support are called strong rules**.**Some competing alternative approaches (other that Apriori) can generate useful rules even with low support values **.**

3.2. C5.0 ALGORITHM

C5.0 is a decision tree tool from Rule quest research that has been installed on storm and runs under UNIX (i.e., it does not run on your PC). C5.0 is a newer, more powerful version of C4.5, a classic decision tree algorithm. C5.0 also provides more features than C4.5, such as support for boosting and cost-sensitive learning. C4.5 is a classic decision tree algorithm.C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1,s_2,...$ of already classified samples. Each sample $s_i = x_1,x_2,...$ is a vector where $x_1,x_2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1,c_2,...$ where $c_1,c_2,...$ represent the class to which each sample belongs. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sub lists. This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

C5.0 also provides more features than C4.5, such as support for boosting and cost-sensitive learning. However, the source code for C5.0 is not available and hence one cannot modify or extend the algorithm. **C4.5** is a classic decision tree algorithm. It has not been modified in many years but still is used for research. It is free and the source code is availableC4.5 made a number of improvements to ID3. Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations. Handling attributes with differing costs. Pruning trees after creation C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with the same nodes.

C5.0 is an extension to this which works more efficiently and accurately. C5.0 has been designed to analyze substantial databases containing thousands to hundreds of thousands of records and tens to hundreds of numeric, time, date, or nominal fields. To maximize interpretability, C5.0 classifiers are expressed as decision trees or sets of if-then rules, forms that are generally easier to understand than neural networks. This algorithm is used in the diagnosing of Hypothyroidism(It is a condition where there is an abnormal low production of the thyroid hormones).The data for this last example comes from an assay screening service related to thyroid function and concern one aspect (hypothyroidism) of thyroid diagnosis. The attributes are a mixture of measured and calculated values and information obtained from the referring physician. There are four classes: negative, primary hypothyroid, secondary hypothyroid, and compensated hypothyroid. Let's show a few examples:

Atrribute	Assay1	Assay2	Assay3
Age	32	63	19
Sex	M	F	M
Sick	T	T	F
Pregnant	N/A	T	N/A
Thyroid	£	£	£
Goitre	£	£	£
Tsh	0.205	108	9
TTH	3.7	4.2	2.3
T3	139	117	123
tumor	£	£	£
Lithium	£	£	£
FTI	104	14	-£

C5.0 processes 2,772 such cases in less than one-tenth of a second, giving seven rules for three of the classes. (The cases for the fourth class were too few in number to justify any rules.)

Rule 1: (31, lift 42.7)
goitre = f
TSH > 6
TT4 <= 37
-> class primary [0.970]

Rule 2: (63/6, lift 39.3)
TSH > 6
FTI <= 65
-> class primary [0.892]

Rule 3: (270/116, lift 10.3)
TSH > 6
-> class compensated [0.570]

Rule 4: (2225/2, lift 1.1)
TSH <= 6
-> class negative [0.999]

Rule 5: (240, lift 1.1)
TT4 > 153
-> class negative [0.996]
Rule 6: (240, lift 1.1)
TT4 > 153
-> class negative [0.996]

Rule 7: (29, lift 1.1)
thyroid surgery = t
-> class negative [0.968]

3.3. COMPARITIVE STUDY (C4.5 VS C5.0)

Both C4.5 and C5.0 can produce classifiers expressed either as decision trees or rule sets. In many applications, rule sets are preferred because they are simpler and easier to understand than decision trees, but C4.5's rule set methods are slow and memory-hungry. C5.0 embodies new algorithms for generating rule sets, and the improvement is dramatic.

- *Accuracy:* The C5.0 rule sets have noticeably lower error rates on unseen cases for the datasets. The C4.5 and C5.0 rule sets have the same predictive accuracy for the seen dataset, but the C5.0 rule set is smaller.
- *Speed:* The times are almost not comparable. For instance, C4.5 required nearly 15 hours to find the lab test for a patient, but C5.0 completed the task in 2.5 minutes.
- *Memory:* C5.0 commonly uses an order of magnitude less memory than C4.5 during rule set construction.

4. CONCLUSION

At present, many data mining methods have been successfully applied to a variety of practical problems in clinical medicine. By combining the data mining methodologies we can solve the problems in emergency department. This study utilized real world data collected from an emergency department of a

hospital and used a new model which is developed combining the Apriori algorithm and a C5.0 algorithm to generate a classification rule base for the classification of chest pain, which can help physicians to make clinical decisions faster and more accurately.

Second, critically ill patients constitute an important proportion of emergency department practice and may remain in the emergency department for significant periods of time. Solutions to emergency department overcrowding may include alternatives for continuing management of critically ill patients. Given the realities of emergency department practice, emergency medicine practitioners should receive training in the continuing management of critically ill patients. So care should be taken in emergency department.

REFERENCES

[1] "A Hybrid Data Mining Method for the Medical Classification of Chest Pain"authored by Sung Ho Ha and Seong Hyeon Joo

[2]"Performance Evaluation of Decision Tree Classifiers on Medical Datasets" authored by D.Lavanya Dr. K.Usha Rani Research Scholar Dept. of Computer Science Sri Padmavathi Mahila Visvavidyalayam

[3] "Medical Domain Knowledge and Associative Classification Rules in Diagnosis"authored by Sung Ho Ha, Kyungpook National University, Korea

[4] "Apriori Algorithm Review for Finals." SE 157B, Spring Semester 2007 Professor Lee By Gaurang Negandhi

[5] http://en.wikipedia.org/wiki/C4.5_algorithm

[6] http://rulequest.com/see5-comparison.html.

[7] "Application of Data Mining Technique for Diagnosis of Posterior Uveal Melanoma" authored by Darius JEGELEVI ˇ CIUS, Ar¯unas LUKOŠEVI ˇ CIUS Institute of Biomedical Engineering, Kaunas University of Technology Alvydas PAUNKSNIS, Valerijus BARZDŽIUKAS Department of Ophthalmology, Institute for Biomedical Research

[8]http://ascelibrary.org/teo/resource/1/jtpedi/v136/i4/p332_s1

[9]http://storm.cis.fordham.edu/~gweiss/resources.html

[10] Jiawei Han and Micheline Kamber Data Mining: Concepts and Techniques, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-

[11]W. T. Lin, S. T. Wang, T. C. Chiang, Y. X. Shi, W. Y. Chen, and H. M. Chen, "Abnormal diagnosis of Emergency Department triage explored with data mining technology: An Emergency Department at a Medical Center in Taiwan taken as an example", *Expert Systems with Applications*, vol. 37, pp. 2733-2741, 2010.

[12]C. Duguary, and F. Chetouane, "Modeling and improving emergency department systems using discrete event simulation," *Simulation*, vol. 83,