International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 21 Issue II, November 2020

A Class Based Approach for Medical Classification of Chest Pain

Manvendra K, Nikita B

***Abstract*— This paper focuses on class based data mining algorithm and their use in medical applications. Data mining techniques have been used in medical research for many years and have been known to be effective. In order to solve such problems as long-waiting time, congestion, and delayed patient care, faced by emergency departments. This study concentrates on building a hybrid methodology, based on using A New Class Based Associative Classification Algorithm which is an advanced and efficient approach than all other association and classification Data Mining algorithms. Applying the association rule into classification can improve the accuracy and obtain some valuable rules and information that cannot be captured by other classification approaches. The class label is taken good advantage in the rule mining step so as to cut down the searching space. The proposed algorithm also synchronize the rule generation and classifier building phases, shrinking the rule mining space when building the classifier to help speed up the rule generation.**

***Keywords*— Data mining, medical decisions, medical domain knowledge, dataset, pruning, rule mining, chest pain.**

I. INTRODUCTION

Emergency departments serve many functions such as 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. To date, the effect of the critically ill patients on the emergency department has not been fully described. Overcrowding usually leads to extremely long wait times, especially for those patients who are not critically ill, which leads to patient dissatisfaction, patient walkouts, and the potential for compromised medical care. Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like " Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease." Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [17]. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Medical history data comprises of a number of tests essential to diagnose a particular disease [8]. Clinical databases are elements of the domain where the procedure of data mining has develop into an inevitable aspect due to the gradual incline of medical and clinical research data. It is possible for the healthcare industries to gain advantage of Data mining by employing the same as an intelligent diagnostic tool. It is possible to acquire knowledge and information concerning a disease from the patient specific stored measurements as far as medical data is concerned. Therefore, data mining has developed into a vital domain in healthcare [6]. It is possible to predict the efficiency of medical treatments by building the data mining applications. Data mining can deliver an assessment of which courses of action prove effective [12] by comparing and evaluating causes, symptoms, and courses of treatments. The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. Working on heart disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process. The researchers in the medical field identify and predict the diseases besides proffering effective care for patients with the aid of data mining techniques. The data mining techniques

have been utilized by a wide variety of works in the literature to diagnose various diseases including: Diabetes, Hepatitis, Cancer, and chest pain. Therefore, the purposes of this study are as follows: using data mining techniques, this study focuses on generating the association rules that help physicians to decide which lab tests patients should be tested by, which can eliminate unnecessary lab tests to classify chest pain diseases and reduce testing time and cost in the emergency department. This study then aims at building a classification scheme that supports to make a complex diagnosis, which can help physicians to formulate clinical decisions more quickly and accurately. The organization of this paper is as follows: Section 2 explains medical data mining and its applications to the clinical decisions. Section 3 illustrates the research methodology used in this study and section 4 illustrates the working of algorithm. Section 5 provides conclusions and future directions.

II. LITERATURE REVIEW

A literature survey showed that there have been several studies on the survivability prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees [7, 8, 9]. In this work, we took the study of Delen et al. [9] as the starting point of our research. In his study, Delen et al. preprocessed the SEER data for to remove redundancies and missing information. The resulting data set had 202,932 records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659) depending on the Survival Time Recode (STR) field. The "survived" class is all records that have a value greater than or equal 60 months in the STR field and the "not survived" class represent the remaining records. After this step, the data mining algorithms are applied on these data sets to predict the dependent field from 16 predictor fields. The results of predicting the survivability were in the range of 93% accuracy. After a careful analysis of the breast cancer data used in [9], we have noticed that the number of "not survived" patients used does not match the number of "not alive" (field VSR) patients in the f irst 60 months of survival time. As a matter of fact, the number of "not survived" patients is expected to be around 20% based on the breast cancer survival statistics of 80% [1]. Khan et al. [6] used decision trees to extract clinical reasoning in the form of medical 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. Yun [7] 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. Ceglowski et al. [10] discovered 'treatment pathways' through mining medical treatment procedures in the emergency department. They found that the workload in the emergency department varies depending on the number of presented patients, and is not affected by the type of procedure carried out. Delphine et al.'s [11] has presented a complementary perspective on the activities of the emergency department for specific patient groups: over 75 year old and under 75 year old patients. She thought once validated, these views would be used as decision support tools for delivering better care to this population. Lin et al. [12] found a way to raise the accuracy of triage through mining abnormal diagnostic practices in the triage. A two-stage cluster analysis (Ward's method, K-means) and a decision tree analysis were performed on 501 abnormal diagnoses done in an emergency department.

III. RESEARCCH METHODOLOGY

A. EMERGENCY DEPARTMENT PROCESS

As patients can present at any time and with any complaint, a key part of the operation of an emergency department is the prioritization of cases based on clinical need. This is usually achieved though the application of triage. Triage is normally the first stage the patient passes through, and most emergency departments have a dedicated area for this to take place, and may have staff dedicated to performing nothing but a triage role. In most departments, this role is fulfilled by a nurse, although dependant on training levels in the country and area, other health care professionals may perform the triage sorting, including paramedics or doctors. Most patients will be assessed and then passed to another area of the department, or another area of the hospital, with their waiting time determined by their clinical need. However, some patients may complete their treatment at the triage stage, for instance if the condition is very minor and can be treated quickly, if only advice is required, or if the emergency department is not a suitable point of care for the patient. Conversely, patients with evidently serious conditions, such as cardiac arrest, will bypass triage altogether and move straight to the appropriate part of the department. The resuscitation area is key in most departments and the most serious patients will be dealt with in this area, and it contains the equipment and staff required for dealing with immediately life threatening illnesses and injuries. Patients whose condition is not immediately life

threatening will be sent to an area suitable to deal with them, and these areas might typically be termed as a *majors* or *minors* area. Children can present particular challenges in treatment and some departments have dedicated pediatrics areas and some departments employ a play therapist whose job is to put children at ease to reduce the anxiety caused by visiting the emergency department, as well as provide distraction therapy for simple procedures. Many hospitals have a separate area for evaluation of psychiatric problems. These are often staffed by psychiatrists and mental health nurses and social workers. There is typically at least one room for people who are actively a risk to themselves or others (e.g. suicidal).Fast decisions on life-and-death cases are critical in hospital emergency rooms. As a result, doctors face great pressures to over test and over treat. The fear of missing something often leads to extra blood tests and imaging scans for what may be harmless chest pains, run-of-the-mill head bumps, and non-threatening stomach aches, with a high cost on the Health Care system.[2]

B.FRAME WORK

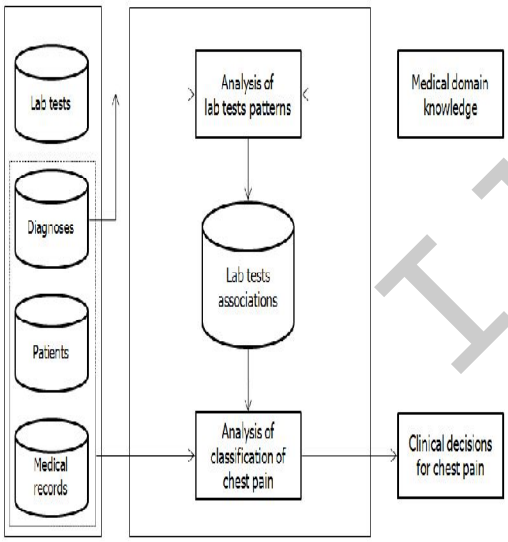
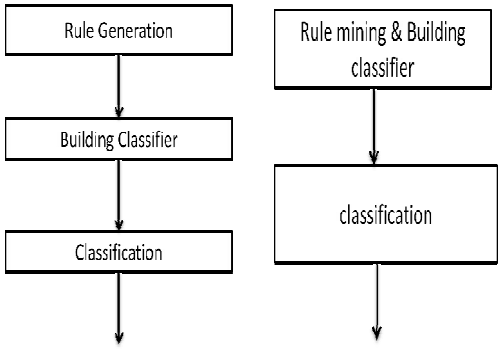


Fig. 1 Research methodology with three stages

Fig. 1 illustrates the research methodology, which consists of three stages. In the first stage, the information of lab tests and diagnoses is collected from the electronic medical data and the association relationships between the lab tests are extracted. The association rules are generated by the Apriori algorithm. The association rule here is an implication of the form $X \Rightarrow Y$, which means 'If a patient takes the lab test X, then he will have the lab test Y.' The rule $X \Rightarrow Y$ has to meet pre-specified minimum support and minimum confidence levels. Domain knowledge about diagnostic tests for a specific disease helps select the lab tests associated with that disease (for example, chest pain). Thus, the second stage selects such lab tests from the generated association rules, which have recorded higher scores (e.g., above 0.9) on support and confidence and have included lab tests of importance mentioned in the domain knowledge. For example, if X is one of the critical tests mentioned in the domain knowledge, then, a test Y will be selected from all the rules with the form of $X \Rightarrow Y$ or $Y \Rightarrow X$, whose support and confidence values are greater than pre-specified minimum support and minimum confidence levels (e.g., 0.9). In the third stage, a classification tree model and its classification rules are generated to classify a chest pain disease, with using the lab tests selected in the second stage and the information of patients (e.g., previous diagnosis and medical records). In this study, a Class based associative classification algorithm, one of associative classification tree models, is chosen to use.

IV. CLASS BASED ASSOCIATIVE CLASSIFICATION ALGORITHM (CACA)

Classification is one of the most important tasks in data mining. Researchers are focusing on designing classification algorithms to build accurate and efficient classifiers for large data sets. Being a new classification method that integrates association rule mining into classification problems, associative classification achieves high classification accuracy, its rules are interpretable and it provides confidence probability when classifying objects which can be used to solve classification problem of uncertainty. Therefore, it becomes a hot theme in recent year. The traditional associative classification algorithms basically have 3 phases: Rule Generation, Building Classifier and Classification as shown in Fig4. Rule Generation employ the association rule mining technique to search for the frequent patterns containing classification rules. Building Classifier phase tries to remove the redundant rules, organize the useful ones in a reasonable order to form the classifier and the unlabeled data will be classified in the third step. However, the drawbacks of associative classification algorithms can be generalized as ,although the associative classification can provide more rules and information, redundant rules may also be included in the classifier which increases the time cost when classifying objects. MCAR[1]determined a redundant rule by checking whether it covers instances in training data set or not .



A.C. Algorithm CACA Algorithm

Fig2: Procedures of A.C. Algorithms

Second, as we know, the rule generation is based on frequent pattern mining in associative classification, when the size of data set grows, the time cost for frequent pattern mining may increase sharply which may be an inherent limitation of associative classification. W. Li, J. Han and J. Pei mine the frequent patterns with FP Growth technique in CMAR which is proved to be very efficient, but extra time should be considered to compute the support and confidence of rules by scanning data set again. In this paper, a class based associative classification algorithm[5] is proposed to solve the difficulty aforementioned. In this algorithm 4 innovations are integrated: 1, use the class based strategic to cut down the searching space of frequent pattern; 2, design a structure call Ordered Rule-Tree to store the rules and their information which may also prepare for the synchronization of the two steps; 3, redefine the compact set so that the compact classifier is unique and not sensitive to the rule reduction; 4, synchronize the rule generation and building classifier phases.

A .Class based rule mining strategy

Given a training data set D with k classes, the principle idea of class based rule mining is to divide the single attribute value set C_{all} for all classes into k smaller ones for every class, that is, to limit the searching in k low dimensional spaces other than a high dimensional one.

B. Ordered Rule Tree Structure (OR-Tree)

To facilitate the synchronization, we design a structure call Ordered Rule Tree under the inspiration of CR-Tree to store and rank rules. It is composed with a tree structure and an ordered list. When a rule $r < a_{i1},a_{i2} .. \ldots\ldots a_{il} , c >$ satisfying the support and confidence thresholds is generated, attribute values $a_{i1},a_{i2},a_{i3},\ldots\ldots..a_{iq}$ will be stored as nodes in this tree according to their frequency in D in descending order. The last node points to an information node storing the rule's information such as class label, support and confidence. Each rule can and only can have one information node. The ordered list is designed to organize all rules in the tree. Each node in the chain points to a certain rule. Nodes pointing to the rules with higher priority are closer to the head node, while those pointing to the rules with lower priority are farther from the head node. When a new non-redundant rule is inserted in the OR-Tree, a new node pointing to this rule will be inserted into a suitable place in the ordered list.

C. Compact Rule Set Redefinition and Pruning Skill

MCAR judge a redundant rule by check whether it cover at least one instance. On one hand, this strategic can not guarantee the removed rule is redundant to the instances n covered by training data set; on the other hand, the reduction should be carry out after all rules are generated and ranked. It is impossible to implement the synchronization with this strategic. However, the definition of compact set and redundant rule in [2], can not ensure the compact classifier is unique and with the same accuracy compared with the original one, which means the classifier and the accuracy changes as the order of rule reduction changes. To overcome these problems, the compact set and redundant rule are redefined in this paper.

1)Redundant Rule:

Given $r_1,r_2,r_3 \in R, r_2$ is redundant if

1.$r_1=<Item_1,c_k>$ and $r_2=<Item_1,c_p>$, but $r_1>r_2$;

2.$r_1=<Item_1,c_k>$, $r_2=<Item_2,c_p>$, $Item_1<item_2$; and $r_1>r_2$

3.$r_1=<Item_1,c_k>$, $r_2=<Item_2,c_k>$ $Item_1<Item_2$ and $r_1>r_2$

4.$r_1=<Item_1,c_k>$, $r_2=<Item_2,c_k>$ $Item_1<Item_2$, $r_2>r_1$ for $r_3=<Item_3,c_p>$, $Item_1<Item_3$, $r_3<r_2<r_1$

2)Compact Rule Set:

For rule set R , if $R` \subset R$ ,any redundant rule $r` \notin R$,and R` is unique, then R`is the compact set of R .

3)Pruning :

For rule $r_i=(item,c_i)$, if $supp(r_i)/conf(r_i). (1-conf(r_i))<minsupp$, stop mining $r_i=(item_k,c_i,)$,$item_k \supset item$.

D. CACA Algorithm

CACA technically combined the rule generation and the building classifier phases together. Once a new rule is generated, the algorithm visits the OR-Tree partially to recognize its redundancy, stores it in the OR-Tree and ranks it in the rule set. Not only can the synchronization simplify the procedure of associative classification but also apply the pruning skill to shrink the rule mining space and raise the efficiency. The algorithm is design as follow:

(1) CACA first scans the training data set D, stores data in form of vertical representation, counts the frequency of every attribute value a_{ij} and arrange a_{ij} in descending order by frequency. The a_{ij} which is failed to satisfy the minsupp is filtered in this step.

(2) For the remaining attribute values a_{ij} in step (1), Intersect $C(a_{ij})$ and $C(c_n)$, n = 1,2,…. k . Add a_{ij} into C_n if $| D(a_{ij}) \cap D(c_n) | >$ minsupp .Thus we have k single attribute value sets $C_1, C_2,\ldots\ldots C_k$.

(3)For class c_n, choose a_{i1j1} C_n in accordance with the order, figure out whether rule r=(a_{i1j1}, c_n,) can satisfy minconf (all the elements in single attribute value sets satisfy support threshold) and its redundancy. If it satisfies the threshold and is not a redundant one, it would be inserted and ranked in the OR-Tree. Check whether it satisfies the condition of pruning skill. If yes, let $C_n = C_n \setminus a_{i1j1}$ and repeat (3), else go on with the recursive procedure of mining more detailed rules.

(4) Take an $a_{i1j2} \in C_n \setminus a_{i1j1}$, $i_1 \neq i_2$with respect to the frequency order. Judge the satisfaction of minsupp for $r = (a_{i1j1}, a_{i2j2}, c_n)$. Any dissatisfaction leads to a new selection of element, that is, select $a_{i3j3} \in C_n \setminus \{a_{i1j1}, a_{i2j2}\}$ and go on with the judgment. Otherwise, if $r = (a_{i1j1}, a_{i2j2}, c_n)$ satisfy them in supp , check the confidence threshold and redundancy as in step (3). Insert the satisfactory rule in the OR-Tree (or modify the OR-tree when an old rule should be replaced by a new one or an old rule become redundant), rank it and check whether the pruning can be applied here. If the pruning can be carried out here, go back to the upper layer of the new rule. When all rules related with c_n and a_{i1j1} is properly recursion. If not, recursively construct Item sets with more attribute values to obtain handled, the recursion is finished. Then let $C_n = C_n \setminus a_{i1j1}$ repeat (3), until $C_n = \varphi$.

(5) Repeat step (3)-(4) until $C_n = \varphi$ n =1,2,…….k .

(6) Classify the unlabeled data by the obtaining classifier.

V. 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 an new model which is developed combining the Apriori algorithm and a Caca 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. According to the characteristic of associative classification, a new class based frequent pattern mining strategic is designed in CACA to cut down the searching space of frequent pattern. OR-Tree structure enables the synchronization of the traditional phases which may not only simplify the associative classification but help to guide the rule generation and speed up the algorithm. And the redefinition of the redundant rule and compact set guarantee the usage of the compact set to help improve the classification efficiency and rule quality won't affect the accuracy of CACA.

REFERENCES

[1] R. E. Fromm, L. R. Gibbs, W. G. McCallum, C. Niziol, J. C. Babcock, A.C. Gueler, and R. L. Levine, "Critical care in the emergency department: a time-based study", *Crit. Care Med.*, vol. 21, pp. 970-976, 1993.

[2] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, pp. 89-109, 2001.

[3] G. Masuda, N. Sakamoto, and R. Yamamoto, "A framework for dynamic evidence based medicine using data mining," In *Proc. 15th IEEE Symposium on Computer-Based Medical Systems*, IEEE press, 2002, pp. 117-122.

[4] M. J. Zaki, "Mining non-redundant association rules," *Data Mining and Knowledge Discovery*, vol. 9, pp. 223-248, 2004.

[5] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann, CA: San Francisco, 2005.

[6] K. H. Butler and S. A. Swencki, "Chest pain: a clinical assessment," *Radiologic Clinics of North America*, vol. 44, pp. 165-179, 2006.

[7] Y. Tan, G. F. Yin, G. B. Li, and J. Y. Chen, "Mining compatibility rules from irregular Chinese traditional medicine database by Apriori algorithm," *Journal of Southwest JiaoTong University*, vol. 15, 2007.

[8] R. Ceglowski, L. Churilov, and J. Wasserthiel, "Combining data mining and discrete event simulation for a value-added view of a hospital emergency department," *Journal of the Operational Research Society*, vol.58, pp. 246-254, 2007.

[9] C. Duguary, and F. Chetouane, "Modeling and improving emergency department systems using discrete event simulation," *Simulation*, vol. 83, pp. 311-320, 2007.

[10] H. Ren, "Clinical diagnosis of chest pain," *Chinese Journal for Clinicians*, vol. 36, 2008.

[11] B. Riccardo and Z. Blaz, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, pp. 81-97, 2008.

[12] Y. P. Yun, "Application and research of data mining based on C4.5 Algorithm," *Master thesis*, Haerbin University of Science and Technology, 2008.

[13] U. Abdullah, J. Ahmad, A. Ahmed, "Analysis of effectiveness of apriori algorithm in medical billing data mining," In *Proc. 4th International Conference on Emerging Technologies*, IEEE press, 2008, pp. 327-331.

[14] R. Delphine, M. Cuggia, A. Arnault, J. Bouget, and P. L. Beux, "Managing an emergency department by analyzing HIS medical data: a focus on elderly patient clinical pathways," *Health Care Management Science*, vol. 11, pp. 139-146, 2008.

[15] F. S. Khan, R. M. Anwer, O. Torgersson, and G. Falkman, "Data mining in oral medicine using decision trees," *International Journal of Biological and Medical Sciences*, vol. 4, pp. 156-161, 2009.

[16] 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.