SIGNIFICANCE OF INTEGRATED TAXONOMY APPROACH IN DIVERSE LIVER CHAUSES

A.S. Aneeshkumar¹, Dr. C. Jothi Venkateswaran²
¹Research Scholar, PG & Research Dept. of Computer Science, Presidency College, Chennai, India, aneesh_kumar777@yahoo.com
²Research Supervisor & Dean, Dept. of Computer Science & Applications, Presidency College, Chennai, India, jothivenkateswaran@yahoo.co.in

ABSTRACT

Liver chaos is an associated factor with other serious diseases like cancer, kidney failure and cardiovascular problems. The preliminary stage of liver dysfunction can be classified as three, which are Hepatitis, Alcoholic Fatty Liver Disease (AFLD) and Non-alcoholic Fatty Liver Disease (NAFLD). In this paper we are applying various integrated classification models of Data Mining for the evaluation of performance.

Keywords: Associative Classification, Classification Based on Association, Classification based on Multiple Association Rules, Classification based on Predictive Association Rules, First Order Inductive Learner, Preprocessing.

I. INTRODUCTION

Many health care organizations having large number of health related data without proper analysis or storage. In a successful organization, it is import to empower the staff and management with data warehousing based on critical thinking and knowledge management tools for strategic decision making [1]. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Regrettably all doctors do not possess expertise in every subspecialty fields and moreover there is a shortage of resource persons in certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together [2].

Determining the early stages of liver disease is better to avoid other consequences. The common liver related problems seen in Indian society are Hepatitis (Viral attack), Alcoholic Fatty Liver Disease (AFLD) and Non-alcoholic Fatty Liver Disease (NAFLD). Hepatitis can be broadly classified as HAV, HBV, HCV, HDV, and HEV. The first three are...
very common in anywhere, but remaining two is seen very rarely and therefore such data are not available for this study. Normally, it is difficult to identify the exact reason of Liver dysfunction in some extend without further examinations and so in this work we are developing a model to predict the accurate type of disease without false assumptions.

Human civilization since the time of immemorial to today, the alcohol is ubiquitous, with constantly changing patterns of alcohol intake around the world [3]. High level of continuous alcoholic consumption will lead to AFLD. Modern life style and lack of exercise are associated with an increased prevalence of diabetes mellitus (DM), Obesity, Hypertension, hyper triglycerides and which are calculated as major reasons of NAFLD. Hepatitis A may be spread primarily through food or water from an infected person and others are through sexual contact with an infected person, injection of drugs, infected women to infants and blood transfer. In some cases toxic drug effects will also be the factor of deviated results of the liver function tests and theses are seen in rare cases. Some hepatitis is considered to have autoimmune hepatitis, on the basis of combination of psychological and histological features, clinical and biochemical features as described by the International Autoimmune Hepatitis Group Report [4].

Radiologic imaging with Ultra-Sonography (US),Computed Tomography (CT) or Magnetic Resonance Imaging, used either singly or in combination, for detection of adequate threshold of fatty infiltrations in the liver. Then in the absence of obvious contraindications, a liver biopsy should be conducted in patients with unexplained sustained abnormalities of the liver biochemistry [5]. The greatest benefit of liver biopsy in unexplained abnormal LFT is to cover progressive liver disease either directly or through prompting a further diagnostic analysis.

II. DATA DESCRIPTION

The collected data for this study from a reputed hospital consists of all symptoms, doctor’s observation and liver biochemistry of the patients, where in addition to Liver Function Test (LFT), Triglycerides report also included for this study. Symptoms of Liver problems are uncertain in most cases. So we neglect such items and the following factors are the attributes for this work.

TABLE 1. Biochemical Markers of LFT

<table>
<thead>
<tr>
<th>Test Items</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilirubin(D)</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Bilirubin(T)</td>
<td>0.3</td>
<td>1.3</td>
</tr>
<tr>
<td>S.G.O.T</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>S.G.P.T.</td>
<td>7</td>
<td>41</td>
</tr>
<tr>
<td>Alkaline Phosphatase</td>
<td>80</td>
<td>290</td>
</tr>
<tr>
<td>Gamma GT</td>
<td>11/7</td>
<td>50/32</td>
</tr>
<tr>
<td>Total Proteins</td>
<td>6.7</td>
<td>8.6</td>
</tr>
<tr>
<td>Albumin</td>
<td>4.1</td>
<td>5.3</td>
</tr>
</tbody>
</table>
### TABLE 2. Other dependent factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBAIC</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Obesity</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>BP</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Alcoholic consumption</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

### III. METHODOLOGY

The combination of knowledge discovery, information retrieval, deductive learning and exploratory data analysis can be called as Data mining [6]. Different algorithms are involved in it to extract models or patterns for particular prediction. The word Association in Mining is finding frequent item sets and discovery of association or correlation among such frequent items in a huge transaction. Another data analysis method called Classification is used to identify some important entities for prediction, where it extracts different models for better grouping of similar data types.

In recent years the data mining researchers proposed a new integrated predictive approach for better accuracy is known as associative classification (AC), which combines associative mining and classification. According to most of the studies, we can see that, it has better accuracy and rule evaluation when compare to other classifications and rule based classification techniques. Because some rules used in AC will not appear in basic classification techniques. AC follows the method of classification in associative rule evaluation and the major application of this is in medical and engineering field. Medical practitioners are always following an inherent approach of combinational symptom and report based diagnosis for disease identification. Normally AC performs its analysis in relational database and each entity consists of attributes like \( a_1, a_2, a_3, \ldots, a_n \) and one of them will consider as Class label or Predictive attribute [7]. Rule generation, rule pruning and classification are the three steps of AC. The first AC algorithm introduced as Classification Based on Association (CBA) and it is based on the Apriori algorithm, which gives Class Associated Rule (CAR). These rules will pruned and then select most appropriate rules from it, for further classification. The CBA will mines all CARs and produce a classification from generated CARs, then it calculate the accuracy of the Association Classifier [8]. Multiple scan features of Apriori produce large number of rules and which maintain large computational time.

Classification based on Multiple Association Rules (CMAR) and Classifications based on Predictive Association Rules (CPAR) are introduced later as the extension of CBA. The CMAR implements FP-growth algorithm instead of Apriori for the generation of frequent item sets [9]. It first generate all the association rules with certain support and confidence thresholds and then selects a small set of rules from them to form a classifier. Then after, it identifies the class label for the best rule with best confidence. When compare
with previous classification it uses multiple rules in prediction, with the help of weighted $X^2$ and so CMAR provides more accuracy in classification.

Classification based on Predictive Association Rules combines the advantages of both associative classification and traditional rule-based classification to attain maximum accuracy. It extends FP growth with interestingness and overlapping relationships in formation of rules. CPAR generates a smaller set of rules, with higher quality and lower redundancy in comparison with other associative classification techniques. So CPAR is much efficient in case of rule generation, prediction and accuracy. More over than that CPAR generates a small set of predictive rules directly from the dataset based on the rule prediction and coverage analysis with lesser time, which is opposite to generating candidate rules like in other Associative Classifications, where datasets contain a large number of rows and columns. So they needs more time to build the model. It follows dynamic programming pattern to avoid repeated calculation in rule generation and select the entire close to the best literals instead of selecting only the best literal only, so the important rules will not lost.

IV. DATA PREPROCESSING

To avoid certain inappropriate, duplicate, unrelated and missing fields it is necessary to implement preprocessing techniques. In our approach, we used dimensionality reduction to avoid patient’s psychodynamic factors and symptoms, because it may show a discrepancy according to the type and variation of Liver disorder. In our previous work, we used Principal Component Analysis (PCA) for this reduction. It searches for $k n$-dimensional orthogonal vectors, which can bitterly represent the data, where $k \leq n$. the original data are thus projected onto a much smaller space, while it results dimensionality reduction. PCA always combines the essence of attributes by creating an alternative, smaller set of variables [10].

V. MODEL BUILDING

In this paper, we build a model to evaluate the performances of four algorithms, where three are incorporated. Three separate and consolidate datasets are used here for this model construction.

5.1. FOIL

First Order Inductive Learner is a sequential covering algorithm which learns first-order logic rules. The tuples of the class for learning rules are called positive tuples, where the remaining is called as negative tuples. The number of positive(negative) tuples covered by the rule $R_1$ is represented as $pos(neg)$ and $pos^l(neg^l)$ be the number of positive(negative) tuples covered by $R^l$. Basic idea of CPAR observed the concept of FOIL [11]. The information gain of rules which provide high accuracy and maximum positive tuples are

$$FOIL\_Gain = \frac{pos^l}{pos^l + neg^l} - \log2 \frac{pos}{pos + neg}$$
which build rules to distinguish positive tuples from negative tuples. In cases of multiple classes, FOIL is applied to each class, where if a rule generates, the positive sample it covers, will be removed until all positive tuples in the data set are covered.

5.2. CBA

CBA uses a heuristic method to construct the classifier, unlike Apriori with the satisfaction of minimum support and minimum confidence level, where the rules are organized according to decreasing precedence based on their confidence and support [12]. If a set of rule with same antecedent, then the rule with the highest confidence is selected to represent the set. When classifying a new tuple, the first rule satisfying the tuple is used to classify it. The classifier also contains a default rule, having lowest precedence, which specifies a default class for any new tuple that is not satisfied by any other rule in the classifier [13]. The set of these rules make a decision set for this classifier.

5.3. CMAR

CMAR uses an enhanced FP-tree that maintains the distribution of class labels among tuples, which satisfying each frequent itemset, and it combine rule generation together with frequent itemset mining in a single step. For an example R1 and R2 are two rules, where antecedent of R1 is more general than that of R2 and \(\text{conf}(R1) \geq \text{conf}(R2)\), then R2 is pruned. That defines the rules with low confidence can be pruned is a more generalized version with higher confidence exists [14].

5.4. CPAR

CPAR relaxes the FOIL application of removing positive samples by allowing the covered tuples to remain there, but it reduces the weight. The resulting rule from each class in this manner will merge to form the classifier rule set. But it produces fewer rules than other CBA and CMAR.

For an example,

\[
\begin{align*}
A_1(1) & \rightarrow A_2(2) \rightarrow A_3(3) \\
A_2(2) & \rightarrow A_3(2) \rightarrow A_4(4) \\
A_3(4) & \rightarrow A_4(1) \rightarrow A_5(3)
\end{align*}
\]

is generated for the following four rules.

\[
\begin{align*}
R1 &= (A_1 = 1, A_2 = 2, A_3 = 3) \\
R2 &= (A_1 = 1, A_3 = 2, A_2 = 4) \\
R3 &= (A_1 = 1, A_2 = 1, A_4 = 2, A_3 = 2) \\
R4 &= (A_1 = 1, A_4 = 1, A_3 = 2, A_2 = 1, A_5 = 3)
\end{align*}
\]
VI. ANALYSIS AND RESULTS

As whispered earlier, because of ambiguity, we applied data reduction technique in preprocessing stage to reduce the dimensions. Then the attributes collected from the patients are shown in Table 1 and 2 for this work in addition to a class diagnosis. In Table 1, we can see the accuracy / Time of the classifier, where the full data set is used as a training set to build the model and again test it with the developed replica. First we applied these algorithms to identify viral infections of liver from the general category. In the second step we applied another set of data for classification, which related to the details of AFLD and the third dataset of NAFLD performs with lesser time. Finally we applied the combination of these three data sets to evaluate the performance. According to these results the average time taken by each method is same in Hepatitis data and the highest accuracy given by CPAR only. In case of AFLD data, four algorithms share a common time of 0.02 seconds with almost same and fine accuracy. The performance of CMAR is somewhat better here. In the next step also CPAR having better performance with 99.0 / 0.0 rate. Finally in case of combined data, when compare to other three FOIL shows poor performance in both cases of accuracy and time. where CPAR completed within 0.02 seconds of time with good result. Figure 1 represents the performance of all these algorithms.

<table>
<thead>
<tr>
<th>DATA TYPES</th>
<th>FOIL</th>
<th>CBA</th>
<th>CMAR</th>
<th>CPAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis</td>
<td>88.26 / 0.04</td>
<td>91.70 / 0.05</td>
<td>91.98 / 0.04</td>
<td>92.0 / 0.04</td>
</tr>
<tr>
<td>AFLD</td>
<td>98.08 / 0.02</td>
<td>98.09 / 0.02</td>
<td>98.40 / 0.02</td>
<td>98.10 / 0.02</td>
</tr>
<tr>
<td>NAFLD</td>
<td>98.38 / 0.01</td>
<td>98.58 / 0.00</td>
<td>98.82 / 0.01</td>
<td>99.0 / 0.0</td>
</tr>
<tr>
<td>Consolidate</td>
<td>94.0 / 0.04</td>
<td>97.27 / 0.04</td>
<td>97.30 / 0.02</td>
<td>97.62 / 0.02</td>
</tr>
</tbody>
</table>

Fig 1. Rule quality performance
VII. CONCLUSION

The observed factors used for this model are very important for liver disease prevalence. Most of the liver problems are presented because of the lifestyle and behaviour of the peoples. In this analysis accuracy shows the quality of the rule. The integration of multiple rules gives more accurate results. As the result the overall performance of CPAR is better than other classifiers. So this approach can be adapted to other application field like medical, insurance, banking, recruitment and psychology for better prediction in the field of their expected results.

REFERENCES

Journal Papers:

Report:

Journal Papers:

Proceedings Papers:
Books:
[10] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Published by Elsevier, second edition – 2006.

Proceedings Papers:

Journal Papers:

Proceedings Papers: