SUBSTANTIATION STUDY FOR DISEASE PREDICTION

Dr. Aneeshkumar A.S.
Assistant Professor & Head, Alpha Arts and Science College, Porur, Chennai, India

***Abstract*—** Statistics in Data Mining is used to identify frequent item sets and its correlation. The broad application of association of Statistics and Data mining in research, market analysis and disease predictions are proved. The identification of influencing factors of the disease is very peculiar work in medical diagnosis. Early detection is always better to reduce disease complexity, time and money expenses also. In this paper we are implementing Statistical Data mining techniques to predict disease.

***Keywords*—Data mining; Naïve bayes classifier; bayesian belief network; FOIL Algorithm; rule- pruning; rule generation; C4.5 decision tree.**

I. INTRODUCTION

In this competitive world, everyone needs to be the best performer and need best performers. In such place the application of data mining is very important to analysis the patterns and predict the knowledge. The most applicable place of data mining is in business and medicine, where most of the data mining techniques are applicable. In business for identify experts, fraud detection, market analysis, sales analysis, performance evaluation and customer group identification are the major fields disease prediction, diagnosis, treatment analysis, research of medicines and other biomedical applications, gene and DNA analysis are some of the fields in medicine. Except this in most of the government sector and public sector also getting the advantages of data mining like forecasting, rural development, polling survey, other project planning, climate and psychological analysis[1].

Classification is one of the task in Data mining, to split the accurate category from a large data base, where statistical method, rule-based methods, decision indexing and integrated approaches are used to get better results. Some of the combined methodologies are Classification based on Association (CBA), Classification based on Multiple Association Rule (CMAR) and Classification based on Predictive Association Rules (CPAR) [2].

Statistical Classifier and rule based classifier are explained in section 3 and 4 respectively

II. STATISTICAL CLASSIFICATION

Statistical classification techniques have a crucial role in the broad area applications of data analysis and prediction. Here we are adopting some major classifiers which are using probability approach and entropy based approach.

A. Probabilistic Classifier

Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data [3]. Bayes, theorem is named after Thomas Bayes, a English clergyman, who did his study in decision and probability theory during 18th century [4]. Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong naive or independence assumptions [5]. That means it is used for predicting class membership probabilities. So the probability model for a classifier is a conditional model defined as

$$\boldsymbol{p(C|F_1, \dots\dots\dots, F_n)}$$

Where, dependent class variable $\boldsymbol{C}$ with number of outcomes conditional on several feature variables $\boldsymbol{F_1, \dots\dots F_n}$. If the number of features is large or when a feature can take on a large number of values, then basing these model on probability tables is infeasible, therefore reformulated model for Bayes' theorem as,

$$\boldsymbol{p(C|F_1, \dots\dots, F_n) = \frac{p(C)p(F_1, \dots\dots, F_n|C)}{p(F_1, \dots\dots, F_n)}}$$

Now the "naive" conditional independence assumptions come into play. Let us assume that each feature F_i is conditionally independent of every other feature $F_j \; for \; j \neq i$. That is

$$\boldsymbol{p(F_i|C, F_j) = p(F_i|C)}$$

Therefore the combined model as

$$\boldsymbol{p(C, F_1, \dots\dots, F_n) = p(C)\, p(F_1|C)\, p(F_2|C) \dots\dots\dots\; p(F_n|C)}$$

$$\boldsymbol{= p(C) \prod_{i=1}^{n} p(F_i|C)}$$

1) Naïve Bayes Classfier

The Naive Bayes classifier combines the above model with a decision rule. One common rule is to pick the hypothesis that is most probable, which is known as the maximum a posteriori (MAP) decision rule. So the classifier is the function here defined as,

$$\boldsymbol{classify(f_1, \dots\dots, f_n) = argmax\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)}$$

2) Bayesian Belief Networks

In Naive Bayes classifier, the value of the attributes is assumed to be conditionally independent of one another, where Beyesian Belief Networks states joint conditional probability distribution. It defines a graphical model of reasonable relationship that provides class conditional independencies to be defined between subsets of variables. So we can say Bayesian Networks combines the essence of probability theory and graph theory which provide a natural tool for dealing with uncertainty and complexity. That means in addition to probability theory, graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.

Nodes of the graph in graphical model represent random variables, and the lack of arcs represents conditional independence assumptions, where directed graphical models, which cannot have directed cycles, also called Bayesian Networks or Belief Networks (BNs), have a more complicated notion of independence.

I. RULE-BASED CLASSIFIER

One of the direct methods to perform classification is generating rules to cover all the cases within the data. Rule $r = (a, c)$, consists of if or $ecedent\ a$, part and the $then$ or $consequent$ part c. The antecedent has a predicate to evaluate the condition of true or false against each tuple in the data base [6]. Mostly rule discovery and evaluation are the popular data mining application especially in the field of medical to identify the hidden disease symptoms [7]. Rule- Base classification carries three parts, which are rule extraction from the data sets, rule evaluation and prediction. Rule extraction means generate candidate rules, which contain all conjunctions of literals that meet the support threshold. Then from these candidates, a subset rule will be selected. These subsets will combine to build best rule for the prediction. Here we are implementing two classifiers known as FOIL Algorithm and C4.5 Decision tree.

A. FOIL

First Order Inductive Learner (FOIL) is a greedy algorithm which proposed by Ross Quinlan, where rules learns positive examples from negative once. The FOIL takes $O(nkm|R|)$ time, where there are n examples, each having k attributes and each attribute having m values on average, and therefore FOIL generates $|R|$ rules [8]. It always repeatedly searches for the current best rule and removes all the positive examples covered by the rule until all the positive examples in the data set are covered [7]. This algorithm use FOIL gain to measure the information gain of rules that cover many positive tuples and having high accuracy, before adding to the previous rule. This is defined as,

$$\boldsymbol{FOIL\_Gain = \frac{pos^I}{pos^I + neg^I} - log2\ \frac{pos}{pos + neg}}$$

Where, pos and neg are the number of positive and negative tuples covered by rule R.

Then to identify the observed differences between the observed distribution and expected distribution of rule covered tuples, a statistical test likelihood ratio can estimate. In this we can see the correlation of the attributes and a class in the significance of correct predictions, which denoted as,

$$\boldsymbol{Likelihood\_Ratio = 2 \sum_{i=1}^{m} f_i\ log\left(\frac{f_i}{e_i}\right)}$$

Where, f_i is the observed frequency of each class i of tuples that satisfy the rule. m is the number of classes and e_i is used to express expected frequency.

1) Rule Pruning

These rules are well performers in case of training data sets, but not that much result oriented in subsequent testing sets. So for managing this problem, the rules will be pruned by removing a attribute sets [2]. FOIL uses an effective pruning method, which is

$$\boldsymbol{FOIL\_Prune(Rule\ R) = \frac{pos - neg}{pos + neg^I}}$$

Here, suppose the pruned value for R is higher, then again R will be pruned.

B. C4.5 Decision Tree

Decision tree is a case indexing technique with inductive approach in case based reasoning. Assigning indexes to cases for further retrieval and comparison is known as case indexing [9]. Decision tree model consists of class-labelled training tuples with a set of rules in a flow-chart manner for dividing a large heterogeneous datasets into a smaller groups with respect to the target variable. These rules can be checked in terms of application by showing them to an expert whether they are meaningful or not [10] and is useful to predict further events, definitions of certain relationships and making most effective decisions by the help of the medical observations [11, 12]. Ross Quinlan in his previous studies in early 1980s, developed decision tree algorithm known as Iterative Dichotomser (ID3) and as an evolution to it in 1993 he presented C4.5 decision tree. This algorithm is used to create univariate decision trees. It first choose attribute for root node

Then create branch for each value of that attribute and split cases according to branches. The splitting ceases when the number of instances to be split is below a certain threshold. Error–based pruning is performed after the growing phase will induce from a training set that incorporates missing values by using corrected gain ratio as splitting criteria. Repeat this process for each branch until all cases in the branch have the same class. In this, the measure of disorder of the data, Entropy is calculated as

$$\boldsymbol{Entropy(\vec{y}) = -\sum_{i=1}^{n} \frac{|y_i|}{|\vec{y}|}\ log\ \frac{|y_i|}{|\vec{y}|}}$$

The conditional entropy is defined as,

$$\boldsymbol{Entropy}(\boldsymbol{i}|\vec{\boldsymbol{y}}) = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{|\boldsymbol{y_i}|}{|\vec{\boldsymbol{y}}|} \boldsymbol{log} \frac{|\boldsymbol{y_i}|}{|\vec{\boldsymbol{y}}|}$$

It iterate all the possible values of entropy and so finally the Gain should be calculated as,

$$\boldsymbol{Entropy\_Gain}(\vec{\boldsymbol{y}}, \boldsymbol{i}) = \boldsymbol{Entropy}(\vec{\boldsymbol{y}} - \boldsymbol{Entropy}(\boldsymbol{i}|\vec{\boldsymbol{y}}))$$

II. DATASETS DESCRIPTION

The collected Liver related data from a reputed hospital in Chennai, having 23 attributes and a class diagnosis. Attributes are the combination of doctor's observation and blood sample report of the supporting tests. Each attribute carries two values, which are 'yes' or 'no'. For this analysis we neglect some rarely seen symptoms. As part of preprocessing technique we will not consider slight variations of liver function tests also. Finally diagnosis having three output values which are hepatitis related diseases, alcoholic fatty liver disease and non-alcoholic fatty liver disease. The collected attributes are given below. For 14 to 20, if there is any serous variations it will represent by 'yes and else 'no'.

TABLE I. List of attributes

NO	ATTRIBUTES	VALUES
1	frequent Alcoholic Consumption	Yes, No
2	Obese	Yes, No
3	Fever	Yes, No
4	Vomiting	Yes, No
5	Abdomen pain	Yes, No
6	Yellowish urinary discharge	Yes, No
7	low appetite	Yes, No
8	disturbance in abdomen	Yes, No
9	pale stools	Yes, No
10	chills	Yes, No
11	rigor	Yes, No
12	head ache	Yes, No
13	acting differently	Yes, No
14	BILIRUBIN (T)	Yes, No
15	BILIRUBIN (D)	Yes, No
16	GAMMA GT	Yes, No
17	ALKALINE PHOSPHATE	Yes, No
18	TOTAL PROTEINS	Yes, No
19	ALBUMIN	Yes, No
20	GLOBULINS	Yes, No
21	Diabetes	Yes, No
22	BP	Yes, No
23	Triglycerides	Yes, No

V. ANALYSIS AND RESUTS

The total data set is used for three type of assessment in each method, which are supply of 100, percentage 10 percentage and 90 percentage of the total dataset as training set where the built model of Naïve bayes classifier as,

TABLE II. BBN structure

#attributes=23 #class index=22

Network structure (nodes followed by parents)

frequent Alcoholic Consumption(4): Diagnosis
Obese(4): Diagnosis
Fever(4): Diagnosis
Vomiting(4): Diagnosis
Abdomen pain(4): Diagnosis
Yellowish urinary discharge(4): Diagnosis
low appetite(4): Diagnosis
disturbance in abdomen(4): Diagnosis
pale stools(4): Diagnosis
chills(4): Diagnosis
rigor(4): Diagnosis
head ache(4): Diagnosis
acting differently(4): Diagnosis
BILIRUBIN (T)(2): Diagnosis
BILIRUBIN (D)(2): Diagnosis
GAMMA GT(2): Diagnosis
ALKALINE PHOSPHATE(2): Diagnosis
TOTAL PROTEINS(2): Diagnosis
ALBUMIN(2): Diagnosis
GLOBULINS(2): Diagnosis
Diabetes(2): Diagnosis
BP(2): Diagnosis
Triglycerides(2): Diagnosis

TABLE III. Generated rule in FOIL algorithm

BILIRUBIN (T) = no AND
Triglycerides = no AND
GAMMA GT = yes: AFL (323.0)
BILIRUBIN (T) = no: NAFL (261.0)

: Hepatitis (145.0)

Number of Rules : 3

TABLE IV. Evaluation of full dataset

Model	Time	Accuracy (%)	Kappa statistics (KS)	Mean absolute error (MAE)	TP Rate	FP Rate	ROC Area
NB	0	100	1	0.0009	1	0	1
NBN	0	100	1	0.0006	1	0	0.799
FOIL	0.02	100	1	0	1	0	1
C4.5	0.02	100	1	0	1	0	1

TABLE V. Evaluation of testing sets

Model	Splitting rate	Time	CCI	ICCI	Accuracy	KS	MAE	TP R	FP R	ROC
NB	10:90	0	496	160	75.6098	0.6225	0.1621	0.756	0.112	0.842
	90:10	0.02	74	0	100	1	0.0005	1	0	1
NBN	10:90	0.03	542	114	82.622	0.7303	0.1173	0.826	0.079	0.749
	90:10	0.02	74	0	100	1	0.0005	1	0	0.794
FOIL	10:90	0.02	656	0	100	1	0	1	0	1
	90:10	0.02	74	0	100	1	0	1	0	1
C4.5	10:90	0.02	656	0	100	1	0	1	0	1
	90:10	0.02	74	0	100	1	0	1	0	1

In Table 4, correctly classified instances (CCI) is 729 and incorrectly classified instances (ICCI) is 0 for all the classifiers.

TABLE VI. Pruned tree of C4.5

```
BILIRUBIN (T) = yes: Hepatitis (145.0)
BILIRUBIN (T) = no
|   Triglycerides = no
|   |   GAMMA GT = no: NAFL (58.0)
|   |   GAMMA GT = yes: AFL (323.0)
|   Triglycerides = yes: NAFL (203.0)

Number of Leaves  :    4

Size of the tree :          7
```

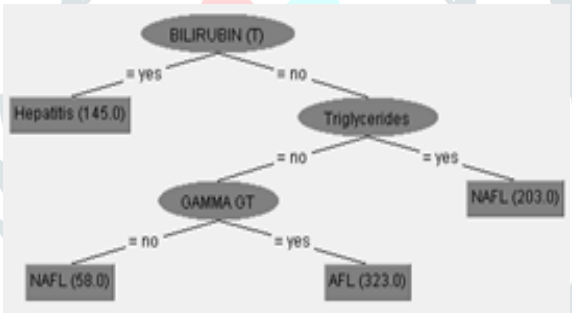

Fig1. Tree formation of C4.5

VI. DISCUSSIONS

In Table 2 we can see the elements of the graph which formed by Bayesian Belief Networks, where each attribute considered as a node and having the value followed from the root node diagnosis. Table 3 shows the final best rule generated by FOIL algorithm after a number of repeated learning and pruning. According to that rule, BILIRUBIN (T), Triglycerides and GAMMA GT having major role in the data set for the classification process of the classes. Another two tables are helpful to compare the classification performance, where Table 4 consists of the recital of full training sets in statistical and rule-based algorithms, that shows all of them having equal performance in case of accuracy, but rule-based classifiers takes more time to build the model when compare to statistical models because of rule learning. Probability methods follow some slight difference between actual value and forecast value. When we are reaching to Table 5, it contains two splits of data, which are in the ratio of 90 and 10 and vice versa. The classification accuracy and correctly classified instances are high in rule-based classifiers. The error rates also more in statistical methods. In this classification Bayesian belief network gives the value of logScore as,

$$\boldsymbol{LogScore\ Bayes}: -12649.926779910485$$
$$\boldsymbol{LogScore\ ENTROPY}: -12469.197706952467$$

According to Table 6, the pruned tree of C4.5 produces rule value. The diagrammatic representation of the same in Figure 1 shows that the size of the generated tree is seven and that having four leaves. The mean square error is insufficient as classification accuracy, as it indicates only the total number of correct classifications. So confusion matrix shows the cross-classification of the predicted class against the true class. The splitting of classification and misclassification into the member of the matrix is the possible method to assign the accuracy differentiations of correct classification and misclassification. The representation of confusion matrix consists TN (True Negative), FP (False Positive), FN (False Negative and TP (True Positive).

TABLE VII. Confusion matrix representation

NB	BBN
a b c <-- classified as 145 0 0 | a = Hepatitis 0 323 0 | b = AFL 0 0 261 | c = NAFL	a b c <-- classified as 145 0 0 | a = Hepatitis 0 323 0 | b = AFL 0 0 261 | c = NAFL
FOIL	**C4.5**
a b c <-- classified as 145 0 0 | a = Hepatitis 0 323 0 | b = AFL 0 0 261 | c = NAFL	a b c <-- classified as 145 0 0 | a = Hepatitis 0 323 0 | b = AFL 0 0 261 | c = NAFL

VII. CONCLUSION

In this paper we build a model to evaluate the performance of two different classification approaches. Probability classifiers are faster than rule-based, but it gives more accuracy in reversible ratio with testing data. That means if the ratio of testing sets is below 50 percentage of the total, it gives more accuracy. When the testing data decrease gradually the accuracy also increase. So we can say that the overall performance of rule-based classifiers is better than statistical classifier.

ACKNOWLEDGEMENT

We express our thanks to the Director Dr.(Capt.) K. J. Jayakumar M.S., M.N.A.M.S., F.A.I.S. and Chief Manager Dr. R. Rajamahendran, B.Sc., M.B.B.S., D.M.C.H., D.H.H.M., P.G.D.H.S.C.(Diab), F.C.D., Sir Ivan Stedeford Hospital, Chennai for providing permission to collect data. We are grateful to the chief Manager for his guidance and also would like to thank other hospital staffs for their valuable suggestions and help throughout this study.

REFERENCES

[1] Duke Hyun Choi, Byeong Seok Ahn and Soung Hie Kim, "Prioritization of association rules in data mining: Multiple criteria decision approach", Expert System with Applications 29(2005) 867-878, ELSEVIER.
[2] W.Li, J.Han and J.Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules. ICDM 01, pg.no:369-376, San Jose, CA, Nov.2001.
[3] A Brief Introduction to Graphical Models and Bayesian Networks, Graphical Models, pg. no: 1-19, http:// www.cs.berkeley.edu/~murphyk/Bayes/bayes.html
[4] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Published by Elsevier, second edition – 2006.
[5] Naive Bayes classifier, Page no:1-8, Wikipedia-Naive-Bayes-Classifier.pdf.
[6] Margaret H.Dunham, "Data Mining Introductory and Advanced Topics", Sixth Impression, 2009
[7] Asha.T, Dr.S. Natarajan and Dr.K.N.B.Murthy, "A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction", IJCA Special Issue on "Artificial Intelligence Techniques-Novel Approaches & Practical Applications", AIT, 2011.
[8] Xiaoxin Yin and Jiawei Han, "CPAR: Classification based on Predictive Association Rules". In Processings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, 2003,pp:369-376.
[9] Lior Rokach and Oded Maimon, "data mining and knowledge discovery handbook", Chapter 9, Decision trees, pg.no:165-192.
[10] "Data Mining and Information Discovery", http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538, March 2000.
[11] SPSS Inc. Answer Tree 2.0 User's Guide, 1998, ISBN 1-56827-254-5.
[12] Bishop. C.M., " Neural Networks for Pattern Recognition" , Clarendon Press, Oxford, 1996.
[13] Aneeshkumar A.S and Jothi Venkateswaran C, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Application, Volume 57, Issue 6, pp.39-42, November 2012.
[14] Aneeshkumar A.S and Jothi Venkateswaran C, "A novel approach for Liver disorder Classification using Data Mining Techniques", Engineering and Scientific International Journal, Volume 2, Issue 1, pp.15-18, March 2015.
[15] Aneeshkumar A.S and Jothi Venkateswaran C, "Significance of Integrated Taxonomy Approach in Diverse Liver Chaoses", International Journal of Computer Engineering & Technology, Volume 3, Issue 3, pp. 265-272, December 2012.