A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction

Asha. T
Department of Information Science & Engg.
Bangalore Institute of Technology
Bangalore, INDIA

Dr. S. Natarajan
Department of Information Science & Engg.
P.E.S. Institute of Technology
Bangalore, INDIA

Dr. K.N.B.Murthy
Department of Information Science & Engg.
P.E.S. Institute of Technology
Bangalore, INDIA

ABSTRACT
Tuberculosis (TB) is a disease caused by bacteria called Mycobacterium tuberculosis. It usually spreads through the air and attacks low immune bodies such as patients with Human Immunodeficiency Virus (HIV). Association Rule Mining (ARM) is one of the most popular approaches in data mining and if used in the medical domain has a great potential to improve disease prediction. This results in large number of descriptive rules. Therefore ARM can be integrated within classification task to generate a single system called as Associative classification(AC) which is a better alternative for predictive analytics. Rule evaluation plays an important role in the rule learning and classification process under Associative classification. Laplace accuracy has been widely used in algorithms such as Classification based on Predictive Association Rules (CPAR) and Predictive Rule Mining (PRM).

In this paper we propose to use CPAR, PRM and First Order Inductive Learner(FOIL) with Statistical test along with Laplace accuracy as rule evaluation measures with different testing modes. We analyze the performance of these methods on TB data with two classes Pulmonary Tuberculosis(PTB) and Retroviral PTB(RPTB) that is those having TB with HIV. This approach helps in the selection of more suitable measure on a particular testing strategy. Results show that CPAR and PRM are almost same and better in accuracy and the number of rules compared to FOIL. Unfortunately when compared in terms of measures the result is same but generation time is less under statistical measure and also rule ordering differs.

General Terms
Data mining, Algorithms, Bioinformatics databases.

Keywords
Tuberculosis, Associative classification, Rule evaluation measures, PTB, RPTB.

1. INTRODUCTION
Rule discovery is one of the most popular data mining techniques, especially in Medicine field because it shows doctor the hidden disease symptoms associated with one another. It is achieved thru Association Rule Mining concept which is defined as follows. Let $\{ t_1, t_2, \ldots t_n\}$ be a set of transactions and let I be a set of items, $I=\{ I_1, I_2, \ldots I_m \}$. An association rule is an implication of the form $X \rightarrow Y$, where X, Y are disjoint subsets of item I and $X \cap Y = \phi$. X is called the *antecedent* and Y is called the consequent of the rule. In general, a set of items such as the antecedent or consequent of a rule is called an *Itemset*. Each *itemset* has an associated measure of statistical significance called *support*. *support(x)=s* is the fraction of the transactions in the database containing X. The rule has a measure of strength called *confidence* defined as the ratio *support(X ∪ Y) / support(X).*

Classification based on association rules has been proved as very competitive [1]. The general idea is to generate a set of association rules with a fixed consequent (involving the class attribute) and then use subsets of these rules to classify new examples. This approach has the advantage of searching a larger portion of the rule version space, since no search heuristics are employed, in contrast to decision tree and traditional classification rule induction. The extra search is done in a controlled manner enabled by the good computational behaviour of association rule discovery algorithms. Another advantage is that the produced rich rule set can be used in a variety of ways without relearning, which can be used to improve the classification accuracy [2].

Rule evaluation measures play an important role in Associative classification. A lot of measures have been proposed in literature in different fields that try to evaluate features of the rules obtained by different types of mining algorithms for association and classification tasks. In this work, we study the predictive power of many of the known evaluation measures such as Statistical test and Laplace Accuracy with CPAR ,PRM and FOIL for the first time on TB data.

India has the world's highest burden of tuberculosis (TB) with million estimated incident cases per year. It also ranks among the world's highest HIV burden with an estimated 2.3 million persons living with HIV/AIDS. Tuberculosis is much more likely to be a fatal disease among HIV-infected persons than persons without HIV infection [3]**.** The microorganisms usually enter the body by inhalation through the lungs. They spread from the initial location in the lungs to other parts of the body via the blood stream. They present a diagnostic dilemma even for physicians with a great deal of experience in this disease.

Previous work[4] involves diagnosing tuberculosis using Artificial Neural Networks(ANN) with multilayer NN and General Regression NN.

2. ASSOCIATIVE CLASSIFICATION
The procedure of associative classification rule mining is not much different from that of general association rule mining.

A typical associative classification system is constructed in two stages: 1) discovering all the event association rules (in which the frequency of occurrences is significant according to some tests); 2) generating classification rules from the association patterns to build a classifier. In the first stage, the learning target is to discover the association rules inherent in a database, but generating frequent itemsets may prove to be quite expensive. The number of rules generated from association rule discovery is quite large. Hence rule pruning is required. Moreover, to avoid the problem of overfitting, proper rule pruning method is to be employed. Ranking of the rules is also important. When a test instance has more than one potentially applicable rules, rule ranking is necessary to prefer one rule over the others. In the second stage, the task is to select a set of relevant association rules discovered to construct a classifier given the predicting attribute.

For example given a rule X -> Y, AC will only consider rules having a target class as the consequent. This means the new integration focuses on a subset of association rules, whose right hand-sides are restricted to the classification class attribute. This type of rule is called Class Association Rules (CARs). While normal association rule allows more than one condition as its consequent and any item from *X* can be the consequent, CARs generated in AC limit the consequent to one fixed target class for each rule and item from *X* are forbid to appear as the class label. In order to perform AC, a classifier will first mine CARs from a given transaction and later select the most predictive rule to perform a classifier [5,6]. AC generates CARs depending on the frequent item generation technique in mining rules. Despite its benefit, AC does propose challenges in its classification performance. The most important thing is to the approach in mining appropriate CARs for the classification and it pruning technology since AC will generate large number of frequent item sets due to its pruning algorithm. Its prominent pitfalls are in its incapability of handling numerical data.

3. BACKGROUND

Different approaches have been proposed for associative classification that has been found to outperform traditional classification algorithms. Some of AC algorithms include Classification based on Association (CBA), Classification based on Multiple Association Rules (CMAR), and CPAR [5, 6]. Generally, AC consists of three main phases, which are rule generation, rule pruning, and classification [7,8].The performance, however, might differ depending on the algorithm employed in any of these three phases. The first AC algorithm was introduced by [1], namely CBA. The algorithm is based on the Apriori association rule algorithm in generating CARs. These rules are later pruned and only one most suitable rule will be used to classify the test set. Essentially, the CBA algorithm performs three tasks. First, it mines all CARs. Second, it produces a classifier from CARs, and finally, it mines normal association rules.

The multiple capabilities in CBA solve a number of problems in traditional classification systems. Since traditional classifiers only generate a small subset of rules that exists in data to form a classifier, the discovered rules may not be interesting. Also, to generate more rules would need the classification system to load the entire database into the main memory. But because CBA generate all rules, the algorithm is more successful in finding interesting rules and the system also allows the data to reside on disk. However, in CBA, the rule generation process might degrade the accuracy of the classifier due to its randomness in selecting the most suitable rule to form the classifier model. CBA inherits Apriori multiple scan features that generates large number of rules, which is costly in terms of large computational time.CMAR is later introduced as the extension to CBA [9]. The CMAR algorithm implements FP-Growth algorithm instead of Apriori in generating its frequent itemset. Next, the subset of matching rules are used to classify a test instance instead of one rule, and this in turn produces better accuracy.

The CMAR algorithm generates and evaluates rules in a similar way as CBA, but uses a more efficient FPtree structure. A major difference is that it uses multiple rules in prediction with associated weights.Nonetheless, when the datasets are large, both rule generation and rule selection in CBA and CMAR are time consuming. The CPAR and other predictive mining algorithms overcome this problem by generating a small set of predictive rules directly from the dataset based on the rule prediction and coverage analysis, as opposed to generating candidate rules. Consequently, CPAR is an improvement to CBA and CMAR [10,11]. The core of CPAR and other predictive mining algorithms is the predictive rule mining capability, whereby after an instance has been correctly covered by a rule, instead of removing it, its weight is decreased by multiplying a factor. This is essentially a greedy approach in rule generation, which is more efficient than generating all candidate rules. CPAR also uses a dynamic programming approach to avoid repeated calculation in rule generation, which in turn more economical. More importantly, CPAR and PRM uses expected accuracy to evaluate rules, and uses the best *k* rules in prediction.

Nada Lavrac et al. demonstrates[12] that many rule evaluation measures developed for predictive knowledge discovery can be adapted to descriptive knowledge discovery tasks.Thabtah F et al.[10] use four associative rule algorithms (CBA, CMAR, CPAR, MCAR) and have compared their performance with reference to accuracy against 12 benchmark classification problems. Kesari verma and O.P.Vyas [13] propose an integrated framework called temporal associative classification with calendar schema. M. Naderi Dehkordi and M. H. Shenassa[14] propose a new classification approach, CLoPAR (Classification based on Predictive Association Rules), which combines the advantages of both associative classification and traditional rule-based classification. NIU Qiang et al.**[15]** propose a new association classification method based on compactness of rules. It extends Apriori Algorithm, which considers the interestingness, importance, overlapping relationships among rules.

4. DATA SOURCE

The medical dataset we are classifying includes 700 real records of patients suffering from TB obtained from a state hospital. The entire dataset is put in one file having many records. Each record corresponds to most relevant information of one patient. Initial queries by doctor as symptoms and some required test details of patients have been considered as main attributes. Totally there are 11 attributes(symptoms) and one class attribute. The symptoms of each patient such as Age, Chroniccough(weeks), Loss of weight, Intermittent fever(days), Night sweats, Sputum, Bloodcough, Chestpain, HIV, Radiographic findings, Wheezing and Class are considered as attributes.

Table 1 shows names of 12 attributes considered along with their Data Types (DT). Type N-indicates numerical and C is categorical.

Table 1. List of Attributes and their Datatypes

No	Name	DT
1	Age	N
2	Chroniccough(weeks)	N
3	WeightLoss	C
4	Intermittentfever	N
5	Nightsweats	C
6	Bloodcough	C
7	Chestpain	C
8	HIV	C
9	Radiographicfindings	C
10	Sputum	C
11	Wheezing	C
12	Class	C

5. METHODOLOGY

5.1 Overview of predictive mining algorithms

FOIL: FOIL (First Or-der Inductive Learner), proposed by Ross Quinlan in 1993 , is a greedy algorithm that learns rules to distinguish positive examples from negative ones. FOIL re-peatedly 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. It makes use of FOIL gain to measure the information gain associated with each attribute before adding to current rule. One reason that FOIL does not achieve as high accuracy is that it generates a very small number of rules. The most time consuming part of FOIL is evaluating every attribute when searching for the one with the highest gain.

PRM: In this section we describe Predictive Rule Mining (PRM) (proposed by Chen,Yin and Huang in 2005), an algorithm which modifies FOIL to achieve higher accuracy and efficiency. Similar to FOIL it uses positive and negative examples but adopts an extra data structure called PNArray to store the information gain to avoid recalculation.In PRM, after an example is correctly covered by a rule, instead of removing it, its weight is decreased by multiplying a factor. This weighted version of FOIL produces more rules and each positive example is usually covered more than once.

In PRM, we generate at least a certain number of rules for each example (de-pending on the weight decay factor). However, these several rules are not necessarily the best rules since they are generated based on greedy algorithm and also these rules are generated from remaining dataset rather than the whole dataset. When selecting attributes during the rule building process, PRM selects only the attribute with the best gain and ignores all others. However there are often a few attributes which are close to the best gain. Since PRM selects only one of them, it may lead to missing some important rules. Both FOIL and PRM algorithms(Yin and Han) are explained below.

ALGORITHM FOIL

Input: Training set D = P ∪ N (P and N are the sets of all positive and negative examples, respectively.)

Output: A set of rules for predicting class labels for examples.

Procedure FOIL

```
rule set R ← Φ
while | P | > 0
    N ' ← N , P ' ← P
    rule r ← empty_rule
   while | N'| > 0 and r.length < max_rule_length
      find the literal p that brings most gain
                according to P' and N'
      append p to r
      remove from P ' all examples not satisfying r
      remove from N ' all examples not satisfying r
  end
  R ← R ∪ { r }
  remove from P all examples satisfying r's body
end
return R
```

ALGORITHM Predictive Rule Mining (PRM)

Input and Output: The same as FOIL Algorithm

Procedure Predictive Rule Mining

```
set the weight of every example to 1
rule set R ← Φ
totalWeight ← TotalWeight (P)
A ← Compute PNArray from D
while TotalWeight (P) > δ . totalWeight
    N ' ← N , P ' ← P , A ' ← A
    rule r ← emptyrule
    while true
       find best literal p according to A '
       if gain(p) < min_gain then break
       append p to r
      for each example t in P'∪N' not satisfying r's
                                                  body
            remove t from P ' or N '
            change A ' according to the removal of t
      end
    end
    R ← R∪{ r }
    for each example t in P satisfying r's body
    t.weight ← α · t.weight
    change A according to the weight decreased
    end
end
return R
```

CPAR: (proposed by Chen, Yin and Huang in 2005) is the modified version of PRM. It stands in the middle between exhaustive and greedy algorithms and combines the advantages of both. CPAR builds rules by adding literals one by one, which is similar to PRM. However, instead of ignoring all literals except the best one, CPAR keeps all close-to-the-best literals during the rule building process. By doing so, CPAR can select more than one literal at the same time and build several rules simultaneously.

5.2 Rule Evaluation Measures

In this section we describe the important measures used in this work. An evaluation metric is needed to determine which conjunct should be added or removed during rule growing process. Accuracy is an obvious choice because it explicitly measures the fraction of training examples classified correctly by the rule. However a potential limitation of accuracy is that it does not take into account the rule's coverage. Hence the following approaches have been discussed to handle the problem.

A Statistical test can be used to prune rules that have poor coverage. The following likelihood ratio statistic is used for this purpose.

$$2 \sum_{i=1}^{k} f_i \log (f_i/e_i)$$

Where k is the number of classes, f_i is the observed frequency of class *i* examples that are covered by the rule and e_i is the expected frequency of a rule that makes random predictions. The statistic has a chi-square distribution with k-1 degrees of freedom. The higher the likelihood ratio is, the more likely that there is a significant difference in the number of correct predictions made by our rule in comparison with a random guessor. That is the performance of our rule is not due to chance. The ratio helps identify rules with insignificant coverage.

Laplace is also an evaluation metric that takes into account the rule coverage. It is given by

$$Laplace = \frac{f_+ + 1}{n+k}$$

Where n is the number of examples covered by the rule, f_+ is the number of positive examples covered by the rule and k is the total number of classes.

5.3 Our Approach

Our proposed approach includes the following steps:

1. Preprocess the numerical and categorical attributes of TB into only binary format.
2. Generate a framework that includes FOIL,PRM and CPAR which provides a choice to statistical measure and Laplace accuracy as rule evaluation measures with 50:50 or 75:25 training and testing modes respectively.
3. Output pruned rules and accuracy.

Pedictive Rule Mining technique requires binary valued input data sets where each record represents an itemset which in turn is a subset of the available set of attributes. These algorithms require that one of the attributes in each record represents a class to which the record is said to belong. Usually, to facilitate identification, either the last or first attribute in each record in the input data represents the class. Thus in our data set the last item in each record represents the class. our TB dataset is first preprocessed by discretizing and normalizing numerical and categorical data items respectively. Discretization techniques can be used to reduce the values of given continuous numerical attributes by dividing the range of attribute into intervals.Normalization converts values associated with categorical data items so that they correspond to unique integer labels. Then we use a framework to select particular algorithm with desired rule metric and testing mode. Each algorithm generates pruned rules and accuracy on different strategy.

6. EXPERIMENTAL RESULTS

Table 2 displays the result of comparing FOIL, PRM and CPAR with statistical and Laplace as evaluation measures on 50:50 and 75:25 training and testing procedures. Algorithm wise CPAR and PRM is better than FOIL in terms of accuracy and number of rules generated. Unfortunately the accuracy and number of rules for each algorithm on different measures will be same but generation time is 0.0secs for statistical than Laplace and also rule ordering differs. Following fig.1 displays snapshot of rule ordering where first column after the rule displays Laplace accuracy and second column displays statistical value. When we select Laplace as rule evaluation measure rules are ordered according to first column value and if we select statistical test as measure, rules are ordered according to second column value.38 and 39 under consequent part in the rule is for PTB and RPTB respectively.

```
Number of rules      = 4
CLASSIFIER
---------- (1) {22} -> {38} 0.99%   230.8
(2) {21} -> {39} 0.97%   215.4
(3) {8 21} -> {39} 0.98%   155.3
(4) {21 36} -> {39} 0.98%   96.25
```

Figure 1: Snapshot of Rule ordering for CPAR with statistical measure

7. CONCLUSIONS

Tuberculosis is an important health concern as it is also associated with AIDS. Retrospective studies of tuberculosis suggest that active tuberculosis accelerates the progression of HIV infection. Recently, methods such as ANN have been intensively used for classification tasks on TB. In this article we apply predictive rule mining techniques such as CPAR, PRM and FOIL with statistical and Laplace as rule evaluation measures for predicting Tuberculosis. CPAR and PRM were better than FOIL and also statistical measure results in less generation time compared to Laplace. Most of the classifier rules help in the best prediction of tuberculosis which even helps doctors in their diagnosis decisions.

8. ACKNOWLEDGEMENTS

Our thanks to KIMS Hospital, Bangalore for providing the valuable real Tuberculosis data and principal Dr. Sudharshan for giving permission to collect data from the Hospital.

Table 2. Comparison of predictive mining algorithms with different measures

Algorithm	Rule Evaluation Measures	Testing mode	Accuracy	Number of Rules generated	Generation Time (secs.)
FOIL	Statistical Test	50:50	98.86%	11	0.0
		75:25	99%	13	0.02
	Laplace	50:50	98.86%	11	0.02
		75:25	99%	13	0.02
PRM	Statistical Test	50:50	99.14%	4	0.0
		75:25	99.29%	5	0.02
	Laplace	50:50	99.14%	4	0.02
		75:25	99.29%	5	0.02
CPAR	Statistical Test	50:50	99.14%	4	0.0
		75:25	99.29%	5	0.02
	Laplace	50:50	99.14%	4	0.02
		75:25	99.29%	5	0.02

9. REFERENCES

[1] B. Liu, W. Hsu, and Y. Ma. Integrating Classification and Association Rule Mining. In KDD '98: Proceedings of the fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 80–86, New York, NY, USA, 1998.ACM Press.

[2] A. Jorge and P. J. Azevedo. An Experiment with Association Rules and Classification: Post-Bagging and Conviction. In A. G. Hoffmann, H. Motoda, and T. Scheffer, editors, Discovery Science, volume 3735 of Lecture Notes in Computer Science, pages 137–149. Springer, 2005.

[3] HIV Sentinel Surveillance and HIV Estimation, 2006. New Delhi, India: National AIDS Control Organization, Ministry of Health and Family Welfare, Government of India. http://www.nacoonline.org/Quick_Links/HIV_Data/ Accessed 06 February, 2008.

[4] Orhan Er, Feyzullah Temurtas and Tantrikulu, A.C. Tuberculosis Disease Diagnosis using Artificial Neural Networks . Journal of Medical Systems, Springer, DOI= 10.1007/s10916-008-9241-x online,2008.

[5] Y. W. C. Chien and Y. L. Chen. Mining Associative Classification Rules with Stock Trading Data – A GA-based Method. Knowledge-based Systems, 23(6):605-614, 2010.

[6] K Kongubol, T. Rakthanmanon, and K.Waiyamai. Using Rule Order Difference Criterion to Decide Whether to Update Class Association Rules. Advances in Intelligent Information and Database Systems, 283:241-252, 2010.

[7] T. D. Do, S.C. Hui, and A. C. M. Fong. Associative Classification with Artificial Immune System. IEEE Transactions on Evolutionary Computation 13(2):217-228,2009.

[8] Z. Tang and Q. Liao. A New Class-based Associative Classification Algorithm. International Journal of Applied Mathematics, 2007.

[9] Li, W., Han, J. and Pei, J. CMAR: Accurate and Efficient Classification based on Multiple Class-Association Rules. In proceedings of IEEE International Conference on Data Mining (ICDM'01). IEEE computer society, Washington, DC, USA, 369–376.2001.

[10] Thabtah, F. A Review of Associative Classification Mining. Journal of Knowl. Eng. Rev., 2(1), 37–65.2007.

[11] F. A. Thabtah, P. Cowling and Y. Peng. Multiple Labels Associative Classification. Knowledge and Information Systems, 9(1):109-129, 2006.

[12] Nada Lavrac, Peter Flach, and Blaz Zupan. Rule Evaluation Measures: A Unifying View. ILP-99, LNAI 1634, pp.174-185,Springer-Verlag Berlin Heidelberg 1999.

[13] Kesari Verma and O. P. Vyas. Classification Based On Calendar Based Temporal Association Rule. ADIT Journal Of Engineering, VOL. 2, NO.1, December 2005.

[14] Naderi Dehkordi, M. H. Shenassa. CLoPAR: Classification based on Predictive Association Rules. In Proceedings of 3rd International IEEE Conference Intelligent Systems. September 2006.

[15] NIU Qiang, XIA Shi-Xiong, ZHANG Lei. Association Classification based on Compactness of Rules. In Proceedings of Second International Workshop on Knowledge Discovery and Data Mining (WKDD). Moscow, 245-247, 2009.

[16] J. Han and M. Kamber. Data mining: Concepts and Techniques: Morgan Kaufmann Pub, 2006.

[17] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition: Morgan Kaufmann Pub, 2005.