16

Data Mining Techniques in the Diagnosis of Tuberculosis

T. Asha[1], S. Natarajan[2] and K. N. B. Murthy[3]
[1]*Department of Information Science & Engineering, Bangalore Institute of Technology,*
[2,3]*Department of Information Science and Engineering, PES Institute of Technology, India*

1. Introduction

Data mining is the knowledge discovery process which helps in extracting interesting patterns from large amount of data. With the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, medical and scientific discovery (J.Han & M.Kamber,2006).

Humans have been manually extracting patterns from data for centuries, but the increasing volume of data in modern times has called for more automated approaches. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1980s). Data mining (DM) is the process of applying these methods to data with the intention of uncovering hidden patterns.

1.1 Data mining process

Generally KDD is an iterative and interactive process involving several steps. This KDD process was chosen (Figure 1) according to UNESCO definition because of its simplicity and comprehensiveness.

1.1.1 Problem identification and definition

The first step is to understand the application domain and to formulate the problem. This step is clearly a prerequisite for extracting useful knowledge and for choosing appropriate data mining methods in the third step according to the application target and the nature of data.

1.1.2 Obtaining and preprocessing data

The second step is to collect and pre-process the data. Today's real-world databases are susceptible to noisy, missing, and inconsistent data due to their typically huge size (often

several gigabytes or more), and their likely origin from multiple, heterogeneous sources. Low quality data will lead to low quality mining results. Data pre-processing is an essential step for knowledge and data mining. Data pre-processing include the data integration, removal of noise or outliers, the treatment of missing data, data transformation and reduction of data etc. This step usually takes the most time needed for the whole KDD process.

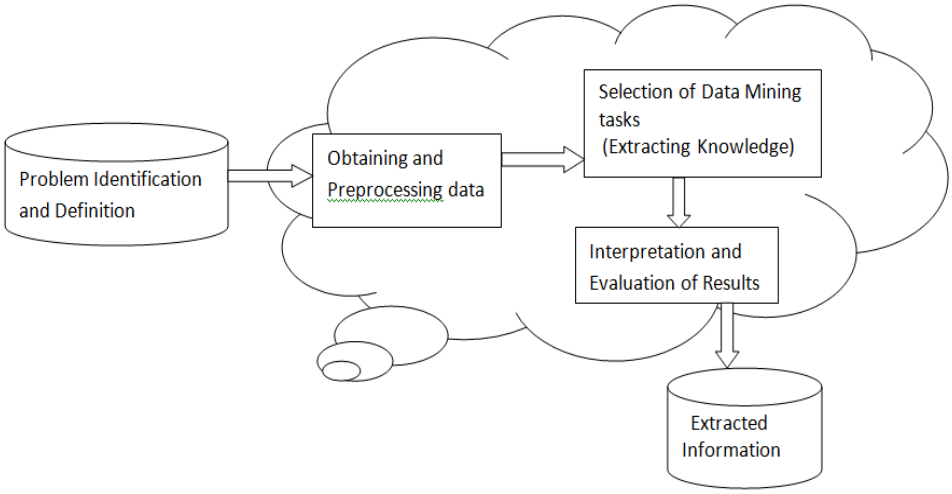


Fig. 1. Structure of Data Mining (KDD) Process

1.1.3 Selection of data mining / knowledge discovery in database

The third step is data mining that extracts patterns and models hidden in data. This is an essential process where intelligent methods are applied in order to extract data patterns. In this step we have to first select data mining tasks and then data mining method. The major classes of data mining methods are predictive modeling such as classification and regression; segmentation (clustering) and association rules which are explained in detail in the next section.

1.1.4 Interpretation and evaluation of results

The fourth step is to interpret (post-process) discovered knowledge, especially the interpretation in terms of description and prediction which is the two primary goals of discovery system in practice. Experiments show that discovered patterns or models from data are not always of interest or direct use, and the KDD process is necessarily iterative with judgement of discovered knowledge. One standard way to evaluate induced rules is to divide the data into two sets, training on the first set and testing on the second. One can repeat this process a number of times with different splits, and then average the results to estimate the rules performance.

1.1.5 Using discovered knowledge

The final step is to put the discovered knowledge in practical use. Putting the results in practical use is certainly the ultimate goal of the knowledge discovery. The information achieved can be used later to explain current or historical phenomenon, predict the future, and help decision-makers make policy from the existed facts (ho, nd).

1.2 Data mining tasks and functionalities

Data Mining functionalities are specifically of two categories: descriptive data mining and predictive data mining. Descriptive methods find human-interpretable patterns that describe the data. Predictive methods perform inference on the current data in order to make predictions (J.Han & M.Kamber, 2006).

The predictive tasks of data mining are:

- Classification - Arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Decision Tree Learning, Nearest neighbor, Naive Bayesian classification and Neural Network.
- Regression -Attempts to find a function which models the data with the least error.

The descriptive tasks of data mining are:

- Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as "market basket analysis".
- Clustering - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.

Data mining finds its applications in various fields. **Web mining -** is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web Usage Mining (WUM), Web Content Mining (WCM) and Web Structure Mining (WSM). It is called **Spatial Data mining** if we apply data mining techniques to spatial data. **Multimedia Data mining**- is the application of data mining techniques to multimedia data such as audio, video, image, graphics etc. **Text mining**- applying data mining techniques on unstructured or semi-structured text data such as news group, email, documents. Bioinformatics and Bio-data analysis on biological data.

Data mining draws ideas from many fields such as Machine learning/Artificial Intelligence, Pattern Recognition, Statistics, and Database Systems. In recent years, data mining has been widely used in the area of genetics, medicine, bioinformatics with its applications applied to biomedical data as facilitated by domain ontologies and mining clinical trial data which is also called medical data mining.

Different types of medical data are now available on the web, where DM algorithms and applications can be applied, helping in easy diagnosis. Efficient and scalable algorithms can

be implemented both in sequential and parallel mode thus improving the performance. Such type of mining is called medical data mining.

1.3 Medical data mining

In recent years, data mining has been widely used in the area of genetics and medicine, called medical data mining. In the past two decades we have witnessed revolutionary changes in biomedical research and bio-technology. There is an explosive growth of bio-medical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations to those identified in genomics and proteomics research. The rapid progress of biotechnology and bio-data analysis methods has led to the emergence and fast growth of a promising new field: Bioinformatics. On the other hand, recent progress in data mining research has led to the developments of numerous efficient and scalable methods for mining interesting patterns and knowledge in large databases, ranging from efficient classification methods to clustering, outlier analysis, frequent, sequential and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools. The question becomes how to bridge the two fields, Data Mining and Bioinformatics, for successful data mining in biomedical data. Especially, we should analyze how data mining may help efficient and effective bio-medical data analysis and outline some research problems that may motivate the further developments of powerful data mining tools for bio data or medical data analysis.

Data mining is a process that involves aggregating raw data stored in a database and analyzing them to identify trends, patterns and anomalies. Medical data mining is an active research area under data mining since medical databases have accumulated large quantities of information about patients and their clinical conditions. Relationships and patterns hidden in this data can provide new medical knowledge as has been proved in a number of medical data mining applications. A Doctor quickly swung into action after a renowned pharmaceutical company in the USA announced in 2001 that it was withdrawing a cholesterol-lowering drug following the deaths of more than 30 people. Using his medical records database, his staff was able to identify all patients taking the cholesterol-lowering drug and notify them within 24 hours of the announcement. What the doctor did is technically known as Data Mining. Very few doctors, however, were able to act on the situation, because they did not have accessible raw data in the electronic format.

Not only does disciplined storage of medical data helps the physicians and healthcare institutions, but it also helps pharmaceutical companies to mine the data to see the trends in diseases. It also helps prioritize product development and clinical trials based on the accurate demands visible from the data that is mined.

Various data mining tasks can be applied on different diseases data set. This helps even the doctor to identify hidden associations between various symptoms. Research has been carried out on gene data, proteonomic data and attributes related to diseases covering even risk factors. Prediction of diseases has also been done on scanned images leading to medical imaging, which is the fastest growing area. Lot of Research has been carried out leading to breast cancer, liver diseases and other types of cancer and also diseases related to heart. There are very few articles related to Tuberculosis.

2. Tuberculosis

Tuberculosis (TB) is a common and often deadly infectious disease caused by mycobacterium; in humans it is mainly *Mycobacterium tuberculosis*. It usually spreads through the air and attacks low immune bodies such as patients with Human Immunodeficiency Virus (HIV). It is a disease which can affect virtually all organs, not sparing even the relatively inaccessible sites. 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. Hence Tuberculosis (TB) is a contagious bacterial disease caused by mycobacterium which affects usually lungs and is often co-infected with HIV/AIDS.

It is a great problem for most developing countries because of the low diagnosis and treatment opportunities. Tuberculosis has the highest mortality level among the diseases caused by a single type of microorganism. Thus, tuberculosis is a great health concern all over the world, and in India as well (wikipedia.org).

Symptoms of TB depend on where in the body the TB bacteria are growing. TB bacteria usually grow in the lungs. TB in the lungs may cause symptoms such as a bad cough that lasts 3 weeks or longer pain in the chest coughing up blood or sputum. Other symptoms of active TB disease are: weakness or fatigue, weight loss, no appetite, chills, fever and sweating at night.

Although common and deadly in the third world, Tuberculosis was almost non-existent in the developed world, but has been making a recent resurgence. Certain drug-resistant strains are emerging and people with immune suppression such as AIDS or poor health are becoming carriers.

2.1 Data set description

The medical dataset we are using includes 700 real records of patients suffering from TB obtained from a city 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 12 attributes (symptoms) and last attribute is considered as class in case of Associative Classification. The symptoms of each patient such as age, chronic cough(weeks), loss of weight, intermittent fever(days), night sweats, Sputum, Bloodcough, chestpain, HIV, radiographic findings, wheezing and TBtype 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.

3. Association Rule Mining

Association Rule Mining (ARM) is an important problem in the rapidly growing field called data mining and knowledge discovery in databases (KDD). The task of association rule mining is to mine a set of highly correlated attributes/features shared among a large number of records in a given database. For example, consider the sales database of a bookstore, where the

records represent customers and the attributes represent books. The mined patterns are the set of books most frequently bought together by the customer. An example could be that, 60% of the people who buy Design and Analysis of Algorithms also buy Data Structure. The store can use this knowledge for promotions, self-placement etc. There are many application areas for association rule mining techniques, which include catalog design, store layout, customer segmentation, telecommunication alarm diagnosis and so on.

No	**Name**	**DT**
1	Age	N
2	chroniccough(weeks)	N
3	weightloss	C
4	intermittentfever(days)	N
5	nightsweats	C
6	Bloodcough	C
7	chestpain	C
8	HIV	C
9	Radiographicfindings	C
10	Sputum	C
11	wheezing	C
12	TBType	C

Table 1. List of Attributes and their Datatypes

3.1 Definition of association rule

Here we give the classical definition of association rules. 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)* (J.Han & M.Kamber, 2006).

Given a set of transactions T, the goal of association rule mining is to find all rules having support ≥ *minsup* threshold and confidence ≥ *minconf* threshold.

Mining Association rule is a Two-step approach:

- Frequent Itemset Generation
  - Generate all itemsets whose support ≥ minsup.
- Rule Generation
  - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset.

Frequent Itemset Generation

The two important algorithms for frequent itemset generation are Apriori algorithm (first proposed by Agrawal, Imielinski and swami VLDB 1994) and Frequent pattern tree growth (FP-Tree) (FPgrowth—Han, Pei & Yin @SIGMOD'00).

Apriori algorithm employs two actions join step and prune step as explained in the following algorithm to find frequent itemsets.

- Apriori principle: It states that if an itemset is frequent, then all of its subsets must also be frequent
- Apriori principle holds due to the following property of the support measure:
  - Support of an itemset never exceeds the support of its subsets
  - $\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$
- This is known as the anti-monotone property of support

Apriori algorithm

- Let k=1
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
  - Generate length (k+1) candidate itemsets from length k frequent itemsets [join step]
  - Prune candidate itemsets containing subsets of length k that are infrequent [prune step]
  - Count the support of each candidate by scanning the DB
  - Eliminate candidates that are infrequent, leaving only those that are frequent

Rule Generation

Once the frequent itemsets from transactions in a database D have been found, it is straightforward to generate strong association rules from them where strong association rules satisfy both minimum support and minimum confidence. This is calculated from the following equation

$$\text{Confidence}(A \rightarrow B) = \text{support\_count}(A \cup B) / \text{support\_count}(A)$$

Based on the above equation association rules can be generated as follows:

- For each frequent itemset l, generate all non empty subsets of l.
- For every nonempty subset s of l, output the rule " s -> (l-s)" if support_count(l) / support_count(s) is greater than or equal to min_conf, where min_conf is the minimum confidence threshold.

Challenges of Apriori

- Multiple scans of transaction database
- Huge number of candidates
- Tedious workload of support counting for candidates

Improving Apriori: general ideas

- Reduce passes of transaction database scans
- Shrink number of candidates
- Facilitate support counting of candidates

Since the processing of the Apriori algorithm requires plenty of time, its computational efficiency is a very important issue. In order to improve the efficiency of Apriori, many researchers have proposed modified association rule-related algorithms.

Advantages of frequent itemset generation and rule generation

- Finding inherent regularities in data
  - What products were often purchased together? — Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

3.2 Tuberculosis association rules

Tuberculosis association rules can be generated by applying data mining ARM technique with the following steps:

- Pre-processing the dataset by discretizing and normalizing
- Generating rules by applying apriori on preprocessed range data

3.2.1 Pre-processing

Incomplete, noisy, and inconsistent data are common among real world databases. Hence it is necessary to preprocess such data before using it. The most common topics under data preprocessing are Data cleaning, Data integration, Data Transformation, Data reduction, Data discretization and automatic generation of concept hierarchies.

Discretization and Normalization are the two data transformation procedures that help in representing the data and their relationships precisely in a tabular format that makes the database easy to understand and operationally efficient. This also reduces data redundancy and enhances performance.

The above TB attributes are normalized and discretized to a suitable binary format. A categorical data field has a value selected from an available list of values. Such data items can be normalized by allocating a unique column number to each possible value. Numerical data fields are discretized by taking values that are within some range defined by minimum and maximum limits. In such cases we can divide the given range into a number of sub-ranges and allocate a unique column number to each sub-range respectively.

Here we give a small example of five patients medical records with five attributes. Table 2 shows original data. Table 3 contains schema of how the attributes are mapped to individual column numbers. Table 4 is the final translated or normalized data.

Age	Chronic cough(weeks)	Weight loss	HIV	Sputum
17	3	Yes	Negative	Yes
13	6	Yes	Negative	Yes
45	6	Null	Negative	Yes
32	Null	Yes	Positive	Null
22	Null	Yes	Positive	Yes

Table 2. Original (raw) Data

Age < 25	Age >= 25	Chronic cough (weeks)< 4	4 <= Chronic cough (weeks) < 8	Chronic cough (weeks) = Null	Weight loss = Yes
1	2	3	4	5	6

Weight loss = Null	HIV = Positive	HIV = Negative	Sputum = Yes	Sputum = Null
7	8	9	10	11

Table 3. Schema Table

In the above tables, note that Age is a numerical attribute and its cut off point is <25 & >=25. Similarly HIV is a categorical attribute where positive value is assigned one number and negative another. The value Null for categorical attribute weightloss is equivalent to No and is assigned a unique number. By using the schema table above we map each tuple in the original data of table 2 to a resulting normalized table shown in table 4. Resulting table has the same number of columns as the original table but filled with unique integer values.

Age	Chronic cough(weeks)	Weight loss	HIV	Sputum
1	3	6	9	10
1	4	6	9	10
2	4	7	9	10
2	5	6	8	11
1	5	6	8	10

Table 4. Normalized Table

3.2.2 TB rules generation

a. **Frequent Itemsets from TB data:** The following figure 2 and 3 shows some of the frequent itemsets generated with 70% support, 90% confidence and 60% support,

[1] {chroniccough(weeks)<39.0} = 692
[2] {weightloss=null} = 550
[3] {chroniccough(weeks)<39.0 weightloss=null} = 544
[4] {intermittentfever(days)<91.25} = 669
[5] {chroniccough(weeks)<39.0 intermittentfever(days)<91.25} = 664
[6] {weightloss=null intermittentfever(days)<91.25} = 530
[7] {chroniccough(weeks)<39.0 weightloss=null intermittentfever(days)<91.25} = 526
[8] {nightsweats=null} = 496
[9] {Bloodcough=null} = 671
[10] {chroniccough(weeks)<39.0 Bloodcough=null} = 663
[11] {weightloss=null Bloodcough=null} = 534
[12] {chroniccough(weeks)<39.0 weightloss=null Bloodcough=null} = 528
[13] {intermittentfever(days)<91.25 Bloodcough=null} = 645
[14] {chestpain=null} = 570
[15 {chroniccough(weeks)<39.0 chestpain=null} = 564
[16] {intermittentfever(days)<91.25 chestpain=null} = 542
[17] {chroniccough(weeks)<39.0 intermittentfever(days)<91.25 chestpain=null} = 539
[18] {Bloodcough=null chestpain=null} = 546

Fig. 2. Frequent Itemsets with 70%support and 90% confidence

80%confidence. The Format of the rule is: [N] {I} = S, where N is a sequential number, I is the item set converted from normalized numerical value to schema text (symptoms) and S the support.

[1] {chroniccough(weeks)<39.0} = 692
[2] {weightloss=null} = 550
[3] {chroniccough(weeks)<39.0 weightloss=null} = 544
[4] {intermittentfever(days)<91.25} = 669
[5] {Bloodcough=null} = 671
[6] {chroniccough(weeks)<39.0 Bloodcough=null} = 663
[7] {weightloss=null Bloodcough=null} = 534
[8] {chestpain=null} = 570
[9] {chroniccough(weeks)<39.0 chestpain=null} = 564
[10] {weightloss=null chestpain=null} = 454
[11] {HIV=Negative} = 465
[12] {chroniccough(weeks)<39.0 HIV=Negative} = 459
[13] {intermittentfever(days)<91.25 HIV=Negative} = 455
[14] {chroniccough(weeks)<39.0 intermittentfever(days)<91.25 HIV=Negative} = 450
[15] {Bloodcough=null HIV=Negative} = 452
[16] {Sputum=yes} = 422
[17] {TBtype=PTB} = 472
[18] {chroniccough(weeks)<39.0 TBtype=PTB} = 466
[19] {intermittentfever(days)<91.25 TBtype=PTB} = 462
[20] {HIV=Negative TBtype=PTB} = 465
[21] {chroniccough(weeks)<39.0 HIV=Negative TBtype=PTB} = 459
[22] {intermittentfever(days)<91.25 HIV=Negative TBtype=PTB} = 455
[23] {intermittentfever(days)<91.25 Bloodcough=null HIV=Negative TBtype=PTB} = 442
[24] {chroniccough(weeks)<39.0 intermittentfever(days)<91.25 Bloodcough=null HIV=Negative TBtype=PTB} = 437

Fig. 3. Frequent Itemsets with 60%support and 80% confidence

b. Discovered TB Association rules

Several medically important association rules are obtained after applying apriori algorithm to the normalized table. It takes the frequent itemsets in figure 2 and 3 and generates rules as shown in figure 4 and 5 respectively. Each association rule shows the relation between one symptom with the other. Data set was first tested by fixing support=70% and confidence=90%.We could get very few association rules, some are listed in figure 4. Rule 7 in figure 4 describes that if weightloss equals null and intermittent fever is less than 91 days implies Bloodcough is null with 97.5% confidence. Most of the rules show the relationship between only few attributes like weightloss, intermittent fever, Bloodcough and chest pain. All the attributes were not shown here. Next with 60% support and 80% confidence we could get large number of association rules, few listed in figure 5 that provides more relationship with many frequent attributes. Rule 1 in figure 5 says if HIV status is negative their TBtype is Pulmonary Tuberculosis (PTB) with 100% confidence. Rule 5 shows the relationship between intermittent fever, Bloodcough, HIV and PTB. Though all the rules

(1) {intermittentfever(days)<91.25 chestpain=null} -> {chroniccough(weeks)<39.0} 99.44
(2) {intermittentfever(days)<91.25 Bloodcough=null chestpain=null} -> {chroniccough(weeks)<39.0} 99.42
(3) {nightsweats=null} -> {chroniccough(weeks)<39.0} 98.99
(4) {chestpain=null} -> {chroniccough(weeks)<39.0} 98.94
(5) {weightloss=null} -> {chroniccough(weeks)<39.0} 98.9
(6) {Bloodcough=null} -> {chroniccough(weeks)<39.0} 98.8
(7) {weightloss=null intermittentfever(days)<91.25} -> {Bloodcough=null} 97.54
(8) {chroniccough(weeks)<39.0 weightloss=null intermittentfever(days)<91.25} -> {Bloodcough=null} 97.52
(9) {chroniccough(weeks)<39.0 weightloss=null Bloodcough=null} -> {intermittentfever(days)<91.25} 97.15
(10) {weightloss=null} -> {Bloodcough=null} 97.09
(11) {Bloodcough=null chestpain=null} -> {chroniccough(weeks)<39.0 intermittentfever(days)<91.25} 95.05
(12) {chestpain=null} -> {chroniccough(weeks)<39.0 Bloodcough=null} 94.73
(13) {chroniccough(weeks)<39.0 chestpain=null} -> {intermittentfever(days)<91.25 Bloodcough=null} 92.02
(14) {chestpain=null} -> {intermittentfever(days)<91.25 Bloodcough=null} 91.57
(15) {chestpain=null} -> {chroniccough(weeks)<39.0 intermittentfever(days)<91.25 Bloodcough=null} 91.05

Fig. 4. Rule generation with 70%support and 90% confidence

(1) {HIV=Negative} -> {TBtype=PTB} 100.0
(2) {chroniccough(weeks)<39.0 HIV=Negative} -> {TBtype=PTB} 100.0
(3) {chroniccough(weeks)<39.0 intermittentfever(days)<91.25 HIV=Negative} -> {TBtype=PTB} 100.0
(4) {Bloodcough=null HIV=Negative} -> {TBtype=PTB} 100.0
(5) {intermittentfever(days)<91.25 Bloodcough=null HIV=Negative} -> {TBtype=PTB} 100.0
(6) {HIV=Negative TBtype=PTB} -> {chroniccough(weeks)<39.0} 98.7
(7) {Bloodcough=null TBtype=PTB} -> {chroniccough(weeks)<39.0} 98.69
(8) {chroniccough(weeks)<39.0 nightsweats=null Bloodcough=null} -> {intermittentfever(days)<91.25} 98.1
(9) {chroniccough(weeks)<39.0 TBtype=PTB} -> {intermittentfever(days)<91.25} 98.06
(10) {TBtype=PTB} -> {chroniccough(weeks)<39.0 intermittentfever(days)<91.25} 96.82
(11) {weightloss=null Bloodcough=null} -> {intermittentfever(days)<91.25} 96.81
(12) {Bloodcough=null TBtype=PTB} -> {chroniccough(weeks)<39.0 intermittentfever(days)<91.25 HIV=Negative} 95.2
(13) {chroniccough(weeks)<39.0 weightloss=null} -> {chestpain=null} 82.53

Fig. 5. Rule generation with 60% support and 80% confidence

may not be interesting to users, only few rules like explained above gives very good description and some hidden relationship may also be found.

We could see from the following output that left side (Antecedent) and right side (consequent) of the rule keep on interchanging repeatedly, which can be pruned by applying some conditions on both antecedent and consequent of a rule.

The format is: (N) ANTECEDENT -> CONSEQUENT CONFIDENCE (%)

4. Associative classification

Association Rule Mining (ARM) as explained in section 3 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.

Classification based on association rules has been proved as very competitive (Liu.B et al., 1998). 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 (Jorge and Azevedo, 2005).

The procedure of associative classification rule mining as shown in figure 6 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 (Chien and Chen, 2010). 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 its 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.

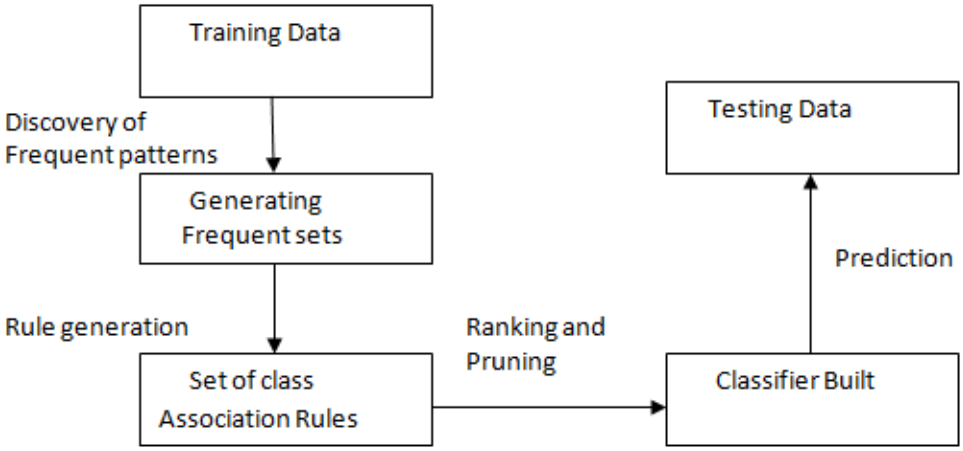


Fig. 6. Associative classification procedure

4.1 Associative Classification Algorithms

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 Classification based on Predictive Association Rules (CPAR-Chien and Chen, 2010). Generally, AC consists of three main phases, which are rule generation, rule

pruning, and classification (Do et al., 2009; Tang and Liao, 2007).The performance, however, might differ depending on the algorithm employed in any of these three phases.

4.1.1 CBA

The first AC algorithm was introduced by (Liu. B et al., 1998), 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**.**

1. Generation of CARs

In CBA, the classification Association rules (CARs) are found iteratively in an apriori algorithm-like fashion. At first, frequent 1-rule itemsets are generated and are pruned. Using this iteratively, other frequent rule itemsets are also found. They are then pruned to get complete set of Classification association rules.

2. Building classifier (Ranking and Pruning Rules)

To prune the rules, CBA uses pessimistic error based pruning method in C4.5. The rule ranking is defined as below:

Given two rules r_i and r_j, $r_i > r_j$ (i.e., r_i precedes r_j or r_i has higher precedence over r_j) if one of the following holds good:

1. The confidence of r_i is greater than that of r_j
2. Their confidences are the same but support of r_i is greater than that of r_j
3. Both the confidences and supports of r_i and r_j are the same, but r_i is generated before r_j

After rule ranking, each training instance is covered by the rule having highest precedence among the rules that can cover the case. Every rule correctly classifies at least one training instance. The rules that do not cover any training instance are removed. The training instances that do not fall into any of the observed classes are added to a default class.

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 we 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.

4.1.2 CMAR

CMAR is later introduced as the extension to CBA (Li et al., 2001). 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**.**

The CMAR algorithm (as described in Li et al., 2001) uses an FP-growth algorithm (Han & Kamber, 2000) to produce a set of CARs and uses CBA method for rule ranking. It prunes rules using high confidence, highly related rules and analyzes the correlation among them using Chi-Squared testing. To test the resulting classifier Li et al. propose the following process.

Given a record r in the test set:

1. Collect all rules that satisfy r, and if consequents of all rules are all identical, or only one rule, classify record according to the consequents.
2. Else group rules according to classifier and determine the combined effect of the rules in each group, the classifier associated with the "strongest group" is then selected. The strength of a group is calculated using a *Weighted Chi Squared* (WCS) measure. Following algorithm shows steps for rule pruning.

Selecting rules based on database:

1. Sort rules in the rank descending order;
2. For each data object in the training data set, set its cover-count to 0;
3. While both the training data set and rule set are not empty, for each rule R in rank descending order, find all data objects matching rule R. If R can correctly classify at least one object then select R and increase the cover-count of those objects matching R by 1. A data object is removed if its cover-count passes coverage threshold δ;

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.

4.1.3 CPAR

CPAR is an improvement to CBA and CMAR (Thabtah et al., 2005; Thabtah, 2007). It is proposed by Chen, Yin and Huang in 2005.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 may choose a number of attributes if those attributes have similar best gain. This is done by first calculating the gain and applying a GAIN_SIMILARITY_RATIO to this. All attributes with gain better than Local Gain Threshold (LGT) are then selected for further processing.

The Local Gain Threshold (LGT) is given by the formula:

$$LGT = bestGain * GAIN\_SIMILARITY\_RATIO$$

Where, GAIN_SIMILARITY_RATIO is a constant whose value is 0.99.

CPAR takes as input a (space separated) binary valued data set R and produces a set of CARs. The resulting classifier comprises a linked-list of rules ordered according to Laplace accuracy. CPAR also uses a dynamic programming approach to avoid repeated calculation in rule generation, which in turn is more economical. More importantly, CPAR selects best *k* rules in prediction.

4.2 Predictive accuracy and rules of associative classifiers

Difference between ARM and AC with reference to results is that the former generates only large number of descriptive rules whereas the latter generate fewer rules along with their performance measure thru accuracy.

CBA generates around 81 rules once it is pruned we get only two rules with an accuracy of 81.14%.

1. { chroniccough(weeks)>23} ->{ TBtype=PTB}
2. { HIV = {Negative}} -> {TBtype=PTB}

CMAR generated about 1091 rules and the pruned output is only 38 rules with an accuracy of 99.1428%. Few are listed below:

1. {HIV = {Negative} } -> { TBtype=PTB}
2. {Bloodcough = {null} HIV = {Negative} } -> {TBtype=PTB}
3. {chroniccough(weeks) <= 22 Bloodcough = {null} HIV = {Negative} } ->TBtype=PTB
4. {HIV = {positive} Sputum = {null}} -> {TBtype=retroviralPTB}
5. {Age>36 HIV = {positive} Sputum = {null}} -> {TBtype=retroviralPTB}
6. {Age>36 chroniccough(weeks)<=22 chestpain={null} HIV={positive} } -> {TBtype=retroviralPTB}

CPAR after pruning could produce only 4 rules with an accuracy of 99.14%.

1. {wheezing = {yes} HIV = {positive}} -> TBtype=retroviralPTB
2. {HIV = {Negative} } -> {TBtype=PTB}
3. {weightloss = {yes} HIV = {positive}} -> {TBtype=retroviralPTB}
4. {HIV = {positive}} -> {TBtype=retroviralPTB}

When compared to both ARM and AC rules, it can be seen that AC rules are smaller and better in description and also CPAR provides better rules compared to all algorithms.

5. Summary

In this chapter two data mining techniques which help in the diagnosis of Tuberculosis have been discussed. Medical databases have accumulated large quantities of information about patients and their clinical conditions and digital era has provided the availability of these information in abundance. Data mining is a knowledge discovery process that helps in

extracting relationships and patterns hidden in this data and can provide a new medical knowledge to doctors in their treatment procedure.

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. It shows doctor the hidden disease symptoms associated with one another. There are many algorithms associated with ARM and the most popular is Apiori. It works in two phases- first is frequent itemset generation where all the items in a database above some minimum specified threshold called support will be generated. Second one is rule generation which generates from the frequent sets, an association rule of the form X->Y based on some minimum confidence. We can say that whenever X appears there is a chance that Y also appears along with it with minimum confidence threshold. These concepts are applied on TB dataset which reveals important association between the symptoms. But this method results in large number of repetitive rules.

Associative classification (AC) is another data mining approach that integrates association rule mining and classification. It uses association rule mining algorithm, such as Apriori or Frequent pattern growth , to generate the complete set of association rules. Then it selects a small set of high quality rules and uses this rule set for prediction. This method results in smaller number rules compared to ARM.

Three important algorithms of AC such as CBA, CMAR and CPAR have been discussed in the chapter. Almost every algorithm contains two major data mining steps, an association rule (AR) mining stage- rules generated here are called as classification association rules (CARs) and a classification stage which uses the mined rules from the first stage directly. The second stage chooses rules with high priority from the CARs to cover training set. The difference between them is based on the priority evaluation of rules which usually depends on the confidence, support, rule length or common quality standard of classification rules. CPAR is better in rule generation compared to others. TB rules and accuracy are compared for every associative classification algorithm

Though the entire rules may not help doctors, few rules may describe the relationship between one symptom with the other and also sometimes it can reveal hidden relationship.

6. References

Agrawal, R., Imielinski, T. and Swami, A. "Mining association rules between sets of items in large databases", In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data (*SIGMOD 93). ACM, New York ,USA ,22(2), 207–216.1993.

Ali. A. El-Solh, M.D., Chiu-Bin Hsiao, M.D., Susan Goodnough RN, et al " Predicting Active pulmonary Tuberculosis using an Artificial Neural network ," CHEST journal 116(4), 968-973,1999.

Antonie, M.-L., Za¨ıane, O. R. and Holte, R.C. "Learning to use a learned model: A two-stage approach to classification", In *Proceedings of the Sixth International Conference*

on Data Mining (ICDM '06, IEEE Computer Society*).* Washington, DC, USA, 33–42, 2006.

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.

Bavani Arunasalam and Sanjay Chawla ,"CCCS: A Top-down Associative Classifier for Imbalanced Class Distribution", In *Proceedings of 12th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'06,* August 20–23 *).* Philadelphia, Pennsylvania, 517-522, 2006.

Carlos Ordonez, Cesar A. Santana , Levien de Braal, "Discovering interesting association rules in medical data," Proc. ACM DMKD 2000 , 78-85 ,2000.

Carlos Ordonez, Edward omiecinski, Cesar A. Santana, et al "Mining Constrained Association Rules to Predict Heart Diseases," Proc. ICDM Nov., 433-440 ,2001.

Carlos Ordonez, "Association Rule Discovery With the Train and Test Approach for Heart Disease prediction'' IEEE Transactions on Information Technology in Biomedicine ,10(2), 334-343,2006.

Chen, T. J., Chou, L. F. and. Hwang, S. J "Application of a data mining technique to analyze coprescription patterns for antacids in Taiwan," *Clin.Ther.*, 25(9), 2453–2463, 2003.

Elisabeth Georgii, Lothar Richter, Ulrich Ruckert and Stefan Kramer "Analyzing Microarray data using quantitative association rules, " Bioinformatics , 21(2) , 123-129, 2005.

Imberman, S., Domanski, B., Thompson, H. "Using dependency/association rules to find indications for computed tomography in a head trauma dataset, " *Artificial Intelligence in Medicine*, Elsevier, 26(1), 55-68,2002.

Keivan Kianmehr, Reda Alhajj, "A class association rule-based classification framework and its application to gene expression data Export," *Artificial Intelligence in Medicine*, Elsevier, 44(1), 7-25,2008.

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

Krzysztof J. Cios, William Moore,G "Uniqueness of medical data mining" *Artificial Intelligence in Medicine*, Elsevier, 26(1), 1–24,2002.

Li, W., Han, J. and Pei, J. ,Cmar: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining(ICDM '01).* IEEE Computer Society, Washington, DC, USA, 369–376. 2001.

Liu, B., Hsu, W. and Ma, Y.,"Integrating Classification and Association Rule Mining", In *ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD '98)*, New York City, NY, 80–86, 1998.

Murat Karabatak, Cevdet Ince, M. "A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases," Expert Systems with Applications:, Elsevier, 36(10), 12500-12505,2009.

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

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.

Orhan Er., Feyzullah Temurtas, Tantrikulu, A.C., "Tuberculosis disease diagnosis using Artificial Neural networks, " Journal of Medical Systems, category: online submission, Springer, DOI 10.1007/s10916-008-9241-x ,2008.

Parameshvyas Laxminarayan, Sergio A. Alvarez, Carolina Ruiz, and Majaz Moonis, " Mining Statistically Significant Associations for Exploratory Analysis of Human Sleep Data," IEEE Transactions on Information Technology in Biomedicine, 10(3) ,440-450, 2006.

Rakesh Agrawal and Ramakrishnan Srikant, " Fast algorithms for mining association rules in large databases ," Proc. VLDB conference Sept 12-15, 487-499 ,1994.

Roddick, J. F., Fule, P. and. Graco, W. J "Exploratory medical knowledge discovery: Experiences and issues," *SIGKDD Explorations*, 5(1), 94–99 ,2003.

Sabeti, M., Sadreddini, M. H., Tahmores Nezhad, J., "EEG Signal Classification Using An Association Rule-Based Classifier", In *Proceedings of IEEE International Conference on Signal Processing and Communications (ICSPC 24-27 November2007)*. Dubai, United Arab Emirates, pp. 620-623, 2007.

Sebban, M. , Mokrousov, I., Rastogi, N. and Sola, C. " A data-mining approach to spacer oligo nucleotide typing of Mycobacterium tuberculosis," Bioinformatics, 18(2), 235-243,2002.

Tamura, Makio , D'haeseleer, Patrik "Microbial genotype-phenotype mapping by class association rule mining," Bioinformatics, 24(13), 1523-1529,2008.

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.

Thabtah, F.,"A review of associative classification mining", *Journal of Knowl. Eng. Rev.*, 2(1), 37–65, 2007.

Thabtah, F., Cowling, P. and Peng, Y., "MCAR: multi-class classification based on association rule approach", In *Proceeding of the Third IEEE International Conference on Computer Systems and Applications*. Cairo, Egypt, 1–7, 2005.

Viet Phan-Luong and Rabah Messouci. "Building Classifiers with Association Rules based on Small Key Itemsets", In *Proceedings of 2nd International Conference on Digital Information Management*. Lyon, 1, 200-205, 2007.

Yanbo, J., Wang , Qin Xin and Frans Coenen. "Novel Rule Weighting Approach in Classification Association Rule Mining", In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops* (ICDMW). 271-276,2007.

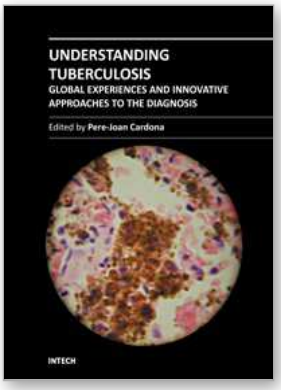
Y. W. C. Chien and Y. L. Chen, "Mining Associative Classification Rules with Stock Trading Data – A GA- based Method", Knowledge-based Systems, 23(6):605-614, 2010.

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

Books

[1] Ian H Witten and Eibe Frank. 2001. Data mining practical machine learning tools and techniques. Morgan Kaufmann publishers.

[2] J. Han and M. Kamber 2006 Data mining: concepts and techniques. Morgan Kaufmann publishers, Sanfrancisco, 47-97.



Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis

Edited by Dr. Pere-Joan Cardona

ISBN 978-953-307-938-7
Hard cover, 552 pages
Publisher InTech
Published online 15, February, 2012
Published in print edition February, 2012

Mycobacterium tuberculosis is a disease that is transmitted through aerosol. This is the reason why it is estimated that a third of humankind is already infected by Mycobacterium tuberculosis. The vast majority of the infected do not know about their status. Mycobacterium tuberculosis is a silent pathogen, causing no symptomatology at all during the infection. In addition, infected people cannot cause further infections. Unfortunately, an estimated 10 per cent of the infected population has the probability to develop the disease, making it very difficult to eradicate. Once in this stage, the bacilli can be transmitted to other persons and the development of clinical symptoms is very progressive. Therefore the diagnosis, especially the discrimination between infection and disease, is a real challenge. In this book, we present the experience of worldwide specialists on the diagnosis, along with its lights and shadows.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821