Application of Data Mining for Exploring Hidden Patterns in Tuberculosis Patients

Farzad Firouzi Jahantigh, Maryam Ostovare
Department of Industrial Engineering
University of Sistan and Baluchestan, Zahedan, Iran
Firouzi@eng.usb.ac.ir, Ostovaremaryam@pgs.usb.ac.ir

Abstract

Having precisely analyzed the data from patients with specific diseases, we could obtain not only the patterns and knowledge of these diseases, but also the specific characteristics of patients. It is usually considered hypotheses in medical studies. Then, the data is prospectively collected to prove or deny this hypothesis, but in many cases, there may be relationships between patient data, which has not even been conjectured about them, and no hypothesis naturally has been considered. Hence, the purpose of this study is to discover the hidden patterns of tuberculosis(TB) patients' datasets. According to various assessment indicators of TB, first, the Entropy-Shannon method was used to identify the most important features. After discovering the existing association rule of data by using the APRIORI technique, the R software was then used to implement these techniques on 548 data of TB patients referred. Following this, the results of the Entropy-Shannon method have identified 18 factors. Also, the APRIORI algorithm was discovered nine association rules between the highest values of lift and the minimum support and confidence value equaled to 0.5 and 0.9, respectively. The results could be considered for further studies, particularly clinical trials as a primary hypothesis or apply to analyze patients' clinical status.

Keywords
Data Mining, Pattern discovery, Feature Selection, Clinical Informatics and Tuberculosis.

1. Introduction

Nowadays, with advances in IT systems, there are the vast sets of patient data and their clinical symptoms. Through detailed data analysis of patients with a particular disease, the patterns and rich knowledge of that disease or even patients own particular features may be achieved. Medical studies consider a hypothesis, and then data is collected prospectively to prove or disprove the hypothesis; however, in most cases, there may be some relations between patient data with no speculation. Data mining techniques can be used to discover such patterns in data. Data mining and knowledge discovery in data is an approach to find the hidden patterns and relationships of data (Wu et al. 2014). Today, data mining has been used in many medical science studies, including Prediction and diagnosis of the diseases (Bhatt et al. 2018), Effectiveness and survival treatments (Rodger 2015), Health Management Services, patient relationship management system (Koh and Tan 2011) and the discovery of hidden patterns in diseases dataset (Sim et al. 2014). New ideas such as knowledge discovery from database including data mining techniques, today, have become more popular and have been transformed into ideal research tools for the researchers. Through them, researchers can identify patterns and relationships between lots of variables and it is made possible for them to predict the results of a disease using information available in databases reserves (Zolbanin et al. 2015).

Before applying data mining techniques for medical data, researchers must understand what kind of data mining algorithms exists and how they function. Generally, data mining algorithms are classified into two categories, predictive (or supervised learning) and descriptive (or unsupervised learning) (Yoo et al. 2012). Prediction data mining infers prediction models from datasets and has the ability to classify the data. Classification is one of the important prediction data mining in which there is a set of records with a field label. In classification, we are looking for a model in order to assign an appropriate label to a record with an unknown label (Buczak et al. 2015). While descriptive data mining, unlike classification, clusters data by measuring the similarity between objects and discovers unknown patterns or relationships in data so that users can readily understand a huge amount of data (Banjari et al. 2015). Association rules which is important technique in this field with exploratory in nature, also determine the dependencies and relationships between data in a database. Discovering such rules has multiple applications in various medicine fields (Sheng et al. 2016). For example, by analyzing patient data and discovering association rules, it can be observed that patients who have sign x have also exhibited sign y or there is a relationship between patients' occupation and their disease that such hypotheses haven't been initially raised (Wei and Scott 2015).

On the other hands, tuberculosis (TB) is one of the most prevalent chronic diseases in the world which takes the lives of numerous people (Sulis et al. 2014). TB is a preventable and treatable disease, with an estimated 9.2 million new cases and two million deaths worldwide (WHO report 2008), it remains one of the leading infectious diseases worldwide (Demay et al. 2012). In many studies, data mining algorithms was conducted on the standardized data to diagnose tuberculosis (Jain and Pardasani 2015; Rastogi and Couvin 2015). However, the study did not discover association rules, but it provided a diagnostic model. On data mining studies which are performed in the diagnosis field, it is noteworthy that such models will not replace practitioners but they playa decision support system (DSS) role for them.

In this study, unlike most other data mining studies, data is collected by researchers merely for doing this study, therefore, clean data is collected for data mining and like many other data mining studies that are conducted retrospectively on existing data collected for other purposes and thus most data preprocessing steps are removed. In addition, in this study, other diseases that TB patients are suffered are also considered in order to obtain possible relationships between a variety of diseases. In this study, we will discover the hidden patterns among TB patients' data using data mining techniques. Researchers can be provided with these patterns and further studies are done to reject or support them since data mining is not provable and it just discovers hidden patterns. On the other hand, these patterns can be provided with practitioners and they can also be applied in patients' status analysis process. Therefore, this study aims to evaluate TB patients' data and discover the potential and hidden relationships among these data.

2. Method

2.1 Data collection

In this study, data from 548 TB patients referred to Masih Daneshvari hospital in Tehran is collected. Data is obtained by referring to the paper records in the archives of hospital, extracting patients' history, the summary of patients' records and referral form. In addition, test results have been extracted individually through the hospital information system.

2.2 Data Discretization

Data is must first be converted to the appropriate format to discover the hidden patterns of data; to do this, all the numeric fields are turned into discrete fields. The equal-width interval discretization method is used to discretize fields that have continuous values (Joiţa 2010).

2.3 Feature selection using entropy shannon

Since there are different and various demographic data and clinical factors of TB patients that influence the creation of rules, the entropy shannon technique used to evaluate and rank the indicators (De Sá et al. 2016). Before implementing this method, field characterizing other diseases to which the patient is affected must be converted to several other fields. Therefore, all diseases in this field are extracted and then new fields with the name of each disease are created based on the number of diseases. For each patient, if he/ she is significant. affected by a disease, "YES" is placed on this field otherwise leave this field empty because if "NO" puts in this field, in a case which disease isn't available, the number of "NO" is much higher than the number of "YES" and more discovered rules will be created on the basis of "NO" values and thus they aren't.

2.4 APRIORI algorithm applied to discover association rules

In the next step, the APRIORI algorithm is applied to discover association rules. This algorithm was originally presented by Agrawal and Srikant (Agarwal and Srikant, 1994) and finds frequent itemsets according to user-defined minimum support and minimum confidence. To better understand the process of implementing the APRIORI algorithm (Yang 2010) is presented in Table 1.

As seen in Table 1, in this algorithm, addition to the database two parameters minimum support and minimum confidence, as the input of the algorithm is given. In the first pass of the algorithm, it constructs the candidate 1-itemsets (C1). The algorithm then generates the frequent 1-itemsets (L1) by pruning some candidate 1-itemsets if their support values are lower than the minimum support and also their confidence values are lower than the minimum confidence. After the algorithm finds all the frequent 1-itemsets, it joins the frequent 1-itemsets with each other to construct the candidate 2-itemsets (C2) and prune some infrequent itemsets from the candidate 2-itemsets to create the frequent 2-itemsets (L2). This process is repeated until no more candidate itemsets can be created (Ye and Chiang 2006).

Table 1- APRIORI algorithm implementation process

```
Input : {Database of transactions, min_sup(minimum support threshold), min_conf(minimum confidence threshold)} ;
$L_1$ = {find frequt1-itemsets} ;
for ($k = 2$; $L_{k-1} \neq \emptyset$; $k++$) {
    $C_k$ = Apriori($L_{k-1}$, min_sup, min_conf) ;
        for each transaction t ∈ Database { #scan Database for counts
            $C_t$ =subset($C_k$ ,t) ; #get the subset of t
                for each candidate $c \in C_t$
            $c.count++$ ;
        }
    $L_k = \{c \in C_k | c.count >$ min_sup, $c.count >$ min_conf} ;
}
Return L=L $\cup_k L_k$ ;
Output : L, find_frequt and rules ;
```

After implementing the APRIORI algorithm, in addition to the two criteria support and confidence to evaluate and select the most appropriate rules, another indicator called lift, which is calculated as the confidence of the rule divided by the support of the right-hand side, can be used to evaluate the resulting rules. Equations (1), (2) and (3) are used to calculate the values of support, confidence and lift respectively.

$$\textbf{Support}\ (\mathbf{X} \to \mathbf{Y}) = \mathrm{P}(\mathrm{X} \cup \mathrm{Y}) = \frac{\mathrm{Count(X)}}{\mathrm{n}} \tag{1}$$

$$\textbf{Confidence}\ (\mathbf{X} \to \mathbf{Y}) = \mathrm{P}(\mathrm{Y} \mid \mathrm{X}) = \frac{\mathrm{Count}(\mathrm{X} \cup \mathrm{Y})}{\mathrm{Count(X)}} \tag{2}$$

$$\textbf{Lift}\ (\mathbf{X} \to \mathbf{Y}) = \frac{\mathrm{P}(\mathrm{X} \cup \mathrm{Y})}{\mathrm{P(X).P(Y)}} \tag{3}$$

Support actually determines possible simultaneous presence of X and Y in X→Y transaction. Confidence indicates the possible presence of Y, if X exists and finally lift represents the probability ratio (Huang et al. 2018). If the lift is equal to 1, X and Y are independent. The more the lift is above 1, the more likely that X and Y will occur together in a transaction because of a relationship between them and not because of a random occurrence. Also if the lift is less than 1, the more likely that X and Y will random occur together in a transaction (Wu et al. 2016). Therefore, to have strong rules, must select the rules with the lift more than 1.

3. Results

In this study 548 data of TB patients referred to Masih Daneshvari hospital from September 2016 to February 2017 with the mean of age 51.96±19.59 were collected. By implementing the entropy shannon method on the factors investigated, 18 features with values greater than 0.0300 were selected. The highest rank belongs to the feature "Chronic cough" with a value of 0.0804. The data on the 18 indicators used to create rules in two terms of numerical and categorical are given in Table 2. For numerical data, mean and standard deviations, and for categorical data, the number and Percentage of the class are considered. Finally, the results of the entropy shannon method can be seen in the final column of table.

Table 2- Data Fields collected from TB patients of Masih Daneshvari hospital

Data Field	N (%)	Mean±SD	Entropy result
Age (year)		51.96±19.59	0.0413
<25	70 (12.77%)	-	-
25-34	93 (16.97%)	-	-
35-44	52 (9.49%)	-	-
45-54	64 (11.69%)	-	-
≥54	269 (49.09%)	-	-
Sex			0.0335
Female	286 (52.19%)	-	-
Male	262 (47.81%)	-	-
Comorbidity			0.0715
No Comorbidity	253 (46.17%)	-	-
Diabetes Mellitus	83 (15.14%)	-	-
Hypertension	59 (10.77%)	-	-
Hepatitis (HBV, HCV)	28 (5.11%)	-	-
HIV	17 (3.10%)	-	-
Ischemic Heart Disease	36 (6.57%)	-	-
Corpulmonale	23 (4.20%)	-	-
Other*	76 (13.87%)	-	-
ESR (mm/hr)	-	61.71±35.94	0.0369
WBC (10*3/μL)	-	8.45±3.07	0.0437
Hb (mg/dl)	-	12.04±1.98	0.0532
Platelet (10*3/μL)	-	300.12±116.04	0.0360
Albumin (g/dl)	-	3.72±0.61	0.0552
BK + Sputum			0.0731
0	136 (24.82%)	-	-
1	158 (28.83%)	-	-
2	84 (15.33%)	-	-
3	170 (31.02%)	-	0.0441
Chronic cough (yes)	487 (88.89%)	-	0.0804
Hemoptysis (yes)	114 (20.80%)	-	0.0326
Weight loss (yes)	406 (74.09%)	-	0.0409
Night sweats (yes)	348 (63.50%)	-	0.0334
Fever (yes)	346 (63.14%)	-	0.0528
History of exposure to TB patient (yes)	99 (18.07%)	-	0.0624
History of smoking (yes)	139 (25.36%)	-	0.0576
History of alcohol (yes)	34 (6.20%)	-	0.0325
Occupation			
Student	54 (9.85%)	-	-
Retired	72 (13.14%)	-	-
Government employee	77 (14.05%)	-	-
Farmer	106 (19.34%)	-	-
Worker	58 (10.85%)	-	-
Other	181 (33.03%)	-	-

*Important diseases with fewer than 10 patients such as Cancer, CRF, CHF, CVA, Asthma, Thyroid, COPD, bronchiectasis.

After that, APRIORI algorithm is applied on data using R software version-3.5.0. A total of more than 200 association rules with the minimum support and confidence value equal to 0.5 and 0.9 respectively, are discovered. Since some of these rules are repetitious and medically meaningless, these rules were shown to a pulmonologist. Finally, table 3 shows 9 discovered rules based on the highest values of support, confidence and lift were extracted as existing and relevant association rules of data.

Table 3- The association rules discovered between the data of TB patients referred to Masih Daneshvari Hospital in Tehran

No	Discovered Pattern	Supp	Conf	Lift
1	Patients who had hemoptysis and their BK + sputum test was 3 also had a chronic cough	0.722	1	1.976
2	Patients who had weight loss and their BK + sputum test was 1 also experienced a chronic cough	0.601	1	1.742
3	Patients who didn't have history of smoking, alcohol and exposure to TB patients, and the number of white blood cells (WBC) is in normal range, didn't have hepatitis.	0.589	0.993	1.436
4	Patients who had chronic cough, weight loss and fever had night sweats	0.525	0.984	1.425
5	Patients who were male in Age 35-44 and had HIV, chronic cough, night sweats and history of smoking also had hepatitis	0.563	0.990	1.256
6	Patients who had night sweats, fever and ESR more than 40 mm/hr also experienced chronic cough	0.515	0.969	1.153
7	Patients who had diabetes mellitus also experienced chronic cough	0.511	1	1.147
8	Patients who were male and had hemoptysis, chronic cough and history of smoking had were worker.	0.555	0.952	1.094
9	Patients who were male and had night sweats also experienced weight loss	0.597	0.92	1.013

4. Discussion

Data mining studies are relatively new in the medical field. In a study conducted by the researchers, there was no similar study to explore the association rules between TB patients' data. Using data mining techniques are particularly useful in medical data since they typically have high amounts and there are many unknown relationships between the causes of diseases and/or demographic characteristics and disease risk sub-factors. Besides statistical methods, numerous medical studies have applied multiple data mining techniques including classification, clustering and association rules (Tseng et al. 2015). In data mining studies, some cases are obtained that there may not have been any thought about them, and after achieving these results, they can be proved or disproved through statistical techniques and developing other studies since data mining acquires only the relationships between data sets in a specific database, and these relationships may not be necessarily established in other data.

This study discovered existing association rules between data of TB patients referred to Masih Daneshvari hospital. After APIRIORI data mining technique and extracting rules, it was found that most rules could be removed via APIRIORI algorithm. APIRIORI algorithm says if the association rule X→Y exists and X is a subset of Z, then it is evident that there is Z→Y and this rule must be removed from discovered rule sets (Prasenna 2012). There were many such rules in the preliminary results of this study. Since APRIORI technique itself wasn't able to remove these rules, these rules were removed manually. Also from a technical viewpoint, if a rule has minimum support and confidence, it will be a meaningful association rule in dataset, but rules initially discovered in this study clearly demonstrated that human user's view, particularly a specialist in the domain of interest is also necessary to refine the rules after technical criteria. So the rules filtered by an expert. After removing such rules, 9 association rules approved in patients' dataset were identified. Association rules discovered in this study are established in this data set but they could be used as hypotheses for other studies and these relationships could be examined in other studies.

The signs of TB depend on which member is attacked by the germs. A TB germ usually involves the lungs and the symptoms of TB include a bad cough for more than three weeks, chest pain, and hemoptysis. Other symptoms may include weakness or fatigue, weight loss, appetite loss, fever, chills and night sweats which sometimes may not be accompanied by none of the above symptoms but it can only be diagnosed with lung accidental X-ray and without any symptoms (World Health Organization 2015). Even though TB is without cough and pulmonary symptoms and only with weight loss, an individual may have some symptoms such as chest pain which is caused and intensified by breathing and coughing. By looking at table 3, it is clear that almost all the main symptoms of TB including hemoptysis, fever, weight loss and night sweats are also associated with chronic cough. Therefore, if there are such symptoms, physicians can diagnose the disease by considering the high possibility of the occurrence of chronic cough in the near future for the patient.

One of the limitations of association rule discovery data mining method is that the discovered rules are established in this dataset and these rules may not be established or changed in another data set of TB patients. Thus, it is better to combine several databases to achieve reliable association rules. Another limitation of the study is the number of patients that were 548 in this study but large amounts of data are typically required for data mining studies. Therefore, in our country, medical and hospital information systems must be strengthened and the storage of patient data and electronic records must also be paid much more attention because if there aren't such storage systems for patients' clinical data, data mining studies with large volumes of data are actually impossible.

The study was conducted on TB patients and since they all had tuberculosis, it was not possible to create data mining models based on classification techniques. It is suggested to carry out a study to establish a classification model for TB diagnosis among those susceptible to the disease by considering TB patient data as well as data from patients who are suspected to have TB but not infected. Also Since the results of a data mining study are difficult to generalize because it is done on a particular data set, it is recommended to conduct systematic review and meta-analysis studies, in the case of quantitative results, on data mining studies that have been carried out around the world about a disease or particular patients so that the overall results could be achieved by analyzing the results of several studies. For instance, a systematic review could be designed and implemented on all data mining studies conducted on tuberculosis.

References

Agarwal, R., and Srikant, R., September. Fast algorithms for mining association rules, In *Proc. of the 20th VLDB Conference*, pp. 487-499, 1994.

Banjari, I., Kenjerić, D., Šolić, K., and L Mandić, M., Cluster analysis as a prediction tool for pregnancy outcomes, *Collegium antropologicum*, vol. 39, no. 1, pp. 247-252, 2015.

Bhatt, A., Dubey, S.K., and Bhatt, A.K., Analytical Study on Cardiovascular Health Issues Prediction Using Decision Model-Based Predictive Analytic Techniques, In Soft Computing: Theories and Applications, pp. 289-299, 2018.

Buczak, A.L., Baugher, B., Guven, E., Ramac-Thomas, L.C., Elbert, Y., Babin, S.M., and Lewis, S.H., Fuzzy association rule mining and classification for the prediction of malaria in South Korea, *BMC medical informatics and decision making*, vol. 15, no.1, pp. 47, 2015.

Demay, C., Liens, B., Burguière, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., and Rastogi, N., SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 755-766, 2012.

De Sá, C.R., Soares, C., and Knobbe, A., Entropy-based discretization methods for ranking data, *Information Sciences*, vol. 329, pp. 921-936, 2016.

Huang, X., Xu, Y., Zhang, S., and Zhang, W., Association Rule Mining for Selecting Proper Students to Take Part in Proper Discipline Competition: A Case Study of Zhejiang University of Finance and Economics, *International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 3, pp.100-113, 2018.

Jain, A., and Pardasani, K.R., Mining fuzzy amino acid associations in peptide sequences of mycobacterium tuberculosis complex (MTBC), *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 4, no. 1, pp. 3, 2015.

Joiţa, D., Unsupervised static discretization methods in data mining, *Titu Maiorescu University, Bucharest, Romania*, 2010.

Koh, H.C., and Tan, G., Data mining applications in healthcare. *Journal of healthcare information management*, vol. 19, no.2, pp. 65, 2011.

Prasenna, P., Kumar, R.K., Ramana, A.R., and Devanbu, A., Network programming and mining classifier for intrusion detection using probability classification. In *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on IEEE*, pp. 204-209, 2012.

Rastogi, N., and Couvin, D., Phylogenetic associations with demographic, epidemiological and drug resistance characteristics of Mycobacterium tuberculosis lineages in the SITVIT2 database: macro-and micro-geographical cleavages and phylogeographical specificities, *International Journal of Mycobacteriology*, vol. 4, pp. 117-118, 2015.

Rodger, J.A., Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive, *Informatics in Medicine Unlocked*, no. 1, pp.17-26, 2015.

Sheng, G., Hou, H., Jiang, X., and Chen, Y., A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model, *IEEE Transactions on Smart Grid*, vol. 9, 2016.

Sim, D.Y., Teh, C.S., and Ismail, A.I., Adaptive Apriori and Weighted Association Rule Mining on Visual Inspected Variables for Predicting Obstructive Sleep Apnea, *Australian Journal of Intelligent Information Processing Systems* vol.14, no.1, 2014.

Sulis, G., Roggi, A., Matteelli, A., and Raviglione, M.C., Tuberculosis: epidemiology and control, *Mediterranean journal of hematology and infectious diseases*, vol. 6, no. 1, 2014.

Tseng, W.T., Chiang, W.F., Liu, S.Y., Roan, J., and Lin, C.N., The application of data mining techniques to oral cancer prognosis, *Journal of medical systems*, vol. 39, no. 5, pp. 59, 2015.

Wei, L., and Scott, J., 2015. Association rule mining in the us vaccine adverse event reporting system (vaers), *Pharmacoepidemiology and drug safety*, vol. 24, no. 9, pp.922-933, 2015.

World Health Organization, *Global tuberculosis control: surveillance, planning, financing*: WHO report 2008, Vol. 393, World Health Organization, 2008.

WHO and World Health Organization, *Guidelines on the management of latent tuberculosis infection*. World Health Organization, 2015.

Wu, X., Zhan, F.B., Zhang, K., and Deng, Q., Application of a two-step cluster analysis and the Apriori algorithm to classify the deformation states of two typical colluvial landslides in the Three Gorges, China, *Environmental Earth Sciences*, vol. 75, no. 2, pp. 146, 2016.

Wu, X., Zhu, X., Wu, G.Q., and Ding, W., Data mining with big data, *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97-107, 2014.

Yang, X.Y., Liu, Z., and Fu, Y., June. MapReduce as a programming model for association rules algorithm on Hadoop. In *Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on IEEE*, pp. 99-102, 2010.

Ye, Y., and Chiang, C.C., August. A parallel apriori algorithm for frequent itemsets mining. In *Software Engineering Research, Management and Applications, 2006, Fourth International Conference on IEEE*, pp. 87-94, 2006.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F., and Hua, L., Data mining in healthcare and biomedicine: a survey of the literature, *Journal of medical systems*, vol. 36, no. 4, pp. 2431-2448, 2012.

Zolbanin, H.M., Delen, D., and Zadeh, A.H., Predicting overall survivability in comorbidity of cancers: A data mining approach, *Decision Support Systems*, vol. 74, pp.150-161, 2015.

Biographies

Farzad Firouzi Jahantigh is an Assistant Professor at the Department of Industrial Engineering at the Sistan and Baluchestan University, Zahedan, Iran. He earned B.S. in Mechanical Engineering from Sistan and Baluchestan University, Zahedan, Masters in Industrial Engineering from Mazandaran University, Babolsar, and PhD in Industrial Engineering from Tarbiat Modares University, Tehran. Dr Firouzi is Senior Lecturer in Industrial Engineering, who is dedicated to teaching, the student experience and his research. He has published journal and conference papers. Although his research specialism is Health Care Engineering, but he is also interested in supply chain management, simulations, quality engineering, management information systems, fuzzy concepts, metaheuristic Algorithms and location & relocation.

Maryam Ostovare is a graduate student in the Department of Industrial Engineering and Optimization Systems from the University of sistan and Baluchestan, Zahedan, Iran. Her primary research focuses on Applied Statistical Process, Optimization and Operations Research topics as applied to Healthcare and Quality Engineering, Bioinformatics, and Metaheuristic Algorithms.