

Review on Data Mining Methods for Tuberculosis Diagnosis

Rusdah*, Edi Winarko**

* Department of Information System, Faculty of Information Technology, Universitas Budi Luhur

** Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Keywords:

Data Mining Methods
Preprocessing Techniques
Clinical Symptoms
Tropical Disease
Review

ABSTRACT

Tuberculosis, which is the oldest human disease with the highest mortality rates among infectious diseases, continues to be the world's attention. Previous methods to diagnose tuberculosis are tuberculin test, Sputum-smear microscopy and chest radiography. Unfortunately, these methods are time consuming and perform poorly. Furthermore, they require varied sensitivity, Mycobacterium tuberculosis bacilli alive, sputum which is difficult to obtain from children, trained personnel to avoid human error, and hence, high cost. Researchers keep developing accurate data mining methods for rapid Tuberculosis diagnosis to reduce the rate of growth of the world population of tuberculosis patients. This paper aims to provide state-of-the-art of data mining methods in diagnosing Tuberculosis using clinical symptoms as input parameters. First, it introduces tuberculosis and current methods used for tuberculosis diagnosis. Then it discusses techniques for preprocessing data and data mining methods for tuberculosis diagnosis currently used. The result shows that the most frequently used variables are sweating at night, more than 3 weeks of cough, fever, weight loss, age, and chest pain respectively. Support Vector Machine and Bayesian Network gave the highest accuracy compared to other methods.

Copyright © 2013 Information Systems International Conference.

All rights reserved.

Corresponding Author:

Rusdah,
Department of Information System, Faculty of Information Technology,
Universitas Budi Luhur,
Jl. Raya Ciledug Petukangan Utara 12260, Jakarta, Indonesia.
Email: rusdah@budiluhur.ac.id

1. INTRODUCTION

World Health Organization (WHO) declared that tuberculosis is a world common disease but needs special attention (emergency situations) [1], [2]. Globally, WHO reported that new cases of TB have been falling for several years and fell at a rate of 2.2% between 2010 and 2011. The TB mortality rate has decreased 41% since 1990 and the world is on track to achieve the global target of a 50% reduction by 2015. Nevertheless, the global burden of TB remains enormous. In 2011, there were an estimated 8.7 million new cases of TB (13% co-infected with HIV) and 1.4 million people died from TB [1]. Tuberculosis is a disease caused by infection of Mycobacterium tuberculosis complex [3]. Persatuan Dokter Paru Indonesia classified Tuberculosis into two types, that is, Pulmonary Tuberculosis and Extra Pulmonary Tuberculosis.

Diagnosis of Tuberculosis (TB) is difficult to do, especially in case of pediatric patients who have a little number of germs or in case of extra pulmonary TB. Several methods have been used to diagnose TB such as clinical symptoms, tuberculin test, sputum-smear microscopy and chest radiography [4]. These methods have several limitations, such as time consuming [5], [6], low performance, difficult to obtain sputum from pediatric patients, requiring live mycobacterium tuberculosis, requiring sophisticated tools that can only be operated by highly skilled medical staff and hence, high cost [7].

In addition, some of the symptoms of TB have in common not only with lung cancer [8], but also with other diseases [9]. It leads to misdiagnosis and death as well [10]. Usually, misdiagnosis can occur because information from the patient or the patient's family is incomplete [11].

To overcome these problems, some researches related to the diagnosis of TB have been done, using sound, images, and variables as input parameters. Some of the researches using sound as an input are [12] which used coughing sound detection algorithm and [6] used lung auscultation software that utilizes lung sound waves to accelerate the process of TB diagnosis with a high degree of accuracy and specificity.

Many researches used image of *Mycobacterium tuberculosis* in tissue as an input to assist pathologists. The method used were Zemike Hybrid Moments and multilayered Perceptron Network [5], feed forward Neural Network [13], the Genetic algorithm - neural network [14], the compact single hidden layer feed forward neural network [15], and the hybridization signal amplification method [16].

In addition, data mining methods to diagnose tuberculosis by using clinical symptoms as input have been used in many studies [17], [18]. This study aims to review state-of-the-art of data mining methods in diagnosing Tuberculosis using clinical symptoms as input parameters.

2. TUBERCULOSIS DATA PREPROCESSING

In order to improve quality of the pattern and information sought, data preprocessing needs to be done before applying data mining techniques [19], [20]. Data preprocessing includes data cleaning (handling missing value and noisy data), data transformation (smoothing, aggregation, generalization, normalization, attribute construction), data integration, data reduction (i.e. data cube aggregation, attribute subset selection, discretization) [19]. By doing data preprocessing, we can provide final training set which is ready to be applied to the data mining techniques [21]. Types of data preprocessing used for preparing tuberculosis data are shown in Table 1.

Missing Value

The raw data tend to be incomplete, noisy, and inconsistent [19]. Data cleaning procedures have to be done to fill the missing values, handle noise in the data outliers, and resolve inconsistent data. The most popular method is to fill in missing values with the most probable value, such as null value [22].

Normalization

Normalization is one way of transformation of the data. An attribute is normalized by making the scale of the attribute value within a small range such as 0.0 to 1.0. This technique is usually used for classification such as neural networks, or for measuring distances such as nearest-neighbor classification and clustering. When using the back propagation neural network algorithm for classification, normalization can accelerate the learning phase. While in the distance-based methods, normalization helps avoid the impression that attributes with large range (e.g., income) is more important than the attribute with a small range (e.g., binary attribute). Some of the normalization methods are min-max normalization, z-score normalization, and normalization by decimal scaling [19]. The min-max method is used to normalize 29 qualitative attributes which contains of 291 Smear Negative Pulmonary Tuberculosis (SNPT) patients data [23].

Feature Selection

Feature subset selection is the process of identifying and reducing the number of attributes which are irrelevant and redundant information as possible [19], [21]. Thereby reducing the data dimensionality will allow the algorithm to run faster and more effectively. Several methods used are information gain, gain ratio, correlation-based feature selection (CSF). Information Gain calculates the information obtained from attribute with respect to the class by using entropy. Gain Ratio method is a form of the normalized Information Gain. Normalization is done by dividing the information gain with the entropy of the attributes with respect to the class, thereby reducing bias. CSF looks for the best attribute which has the highest correlation with the class attribute but the lowest correlation between each attribute [20]. Information Gain method is applied to handle tuberculosis patient data [24], [18]. The study used 20 of 30 attributes for diagnosing tuberculosis. While [23] used single factor logistic regression to generate 29 of 49 attributes from SNPT patients data as Tuberculosis disease diagnostic variable.

Discretization

Discretization process is needed in classification algorithms. Discretization is to divide the continuous value of an attribute into intervals. Discretization algorithm consists of supervised algorithms that discretize attributes by using class information and unsupervised algorithms that discretize attributes without using class information [19], [20]. Equal size discretization, including in unsupervised direct method is the simplest one [21]. The method calculates the maximum and minimum value of discretized attributes and partitions a range into k intervals of equal size. Other unsupervised methods are equal frequency. This method calculates the number of existing values in an attribute that would be discretized and divides it into several intervals with the same number of instances [21]. Discretization methods used in the numerical data are binning, histogram analysis, entropy-based discretization, c2-merging, cluster analysis, and discretization by intuitive partitioning [19]. Some discretization methods used by researchers to handle tuberculosis data were rough set applications [17], Equal size discretization with min-max limit [25] and MDL-based discretization method [18].

Table 1. Preprocessing Tuberculosis Data

Author(s)	Datasets Used	Number of Variable used	Variables Used	Preprocessing	TB Types		
					PTB*	Retroviral PTB	NM*
Bakar & Febriyani, 2007	From Unit of Health Service Mandau, Riau 14 attributes 187 of 233 rec	8	Sex, Age, Weight, Fever, Night Sweat, Cough > 3 weeks, Blood Phlegm, Sputum Test	Discretization: rough set application			√
Benfu et al., 2009	From TB dispensaries and hospital in Jining City 291 rec of patients SNPT 5 major & 41 minor characteristics	29	Gender, Age, Marriage, Occupation, Vaccinated, Chronic disease, Duration of cough, Expectoration, Duration of Expectoration, Night Sweat, Duration of Night Sweat, Low-grade fever, Low-grade fever in the afternoon, debility, no appetite, weight loss, degree of loss, chest pain, chest distress, duration of distress, lesion area, lesion extent, conglutination, fluidly, cavity, Red Blood Cells, White Blood Cells, <i>Erythrocyte Sedimentation Rate</i>	Feature selection : Single factor logistic regression Normalization: Min-Max	√		
Asha et al., 2011	From city hospital 700 rec	11	Age, Chronic cough (weeks), loss of weight, blood cough, intermittent fever (days), chest pain, HIV, radiographic findings, wheezing, night sweat, sputum	Missing value: replace with Null	√	√	
Uçar & Karahoca, 2011	From clinic 667 rec 30 attributes	20	Age Group, Weight, Exhaustion, Unwillingness to work, Loss of Appetite, Loss in weight, Sweating at night, Hemoptysis, Fever, Sedimentation, PPD, Erythrocyte, Hematocrit, Hemoglobin, Leucocyte, Number of leucocyte type, Active specific lung lesion, Calcific tissue, Cavity, Pneumonic infiltration	Feature selection : information gain ranking filter using WEKA function			√
Ali et al, 2012	From Epi-Lab Sudan 54 Attributes, 714 Records	54	NM	NM	√	√	
Ansari et al., 2012	70 rec training data	5	Age, Immune System, Alcohol Intake, Economic Status, International Connection				√
Asha et al., 2012	From city hospital 700 rec	11	Age, Chronic cough (weeks), loss of weight, blood cough, intermittent fever (days), chest pain, HIV, radiographic findings, wheezing, night sweat, sputum	Discretization: Min-Max Limit	√	√	
Uçar et al., 2012	From Private Health Clinic in Istanbul 250 records data sets	20	Active specific lung lesion, Calcific tissue, Number of leucocyte types, Weight, Fever, Age Group, PPD, Sweating at nights, Leucocyte, Loss in weight, Hemoptysis, Cavity, Sedimentation, Loss of appetite, Pneumonic infiltration, Exhaustion, Unwillingness for work, Hemoglobin, Hematocrit, Erythrocyte	Feature selection : information gain ranking filter using WEKA function Discretization MDL-based discretization method	√		

*) PTB: Pulmonary Tuberculosis; NM: Not Mentioned

3. VARIABLES USED FOR TUBERCULOSIS DIAGNOSIS

Different types and numbers of variables have been used to diagnose tuberculosis. From table 1, we can see that the most widely used variables in the diagnosis of tuberculosis are age, sweating at night, more than 3 weeks of cough, fever, weight loss, age, and pain in the chest respectively.

4. TYPES OF TUBERCULOSIS

From Table 1, it can be seen that the most commonly used type of TB is pulmonary TB as well as HIV-infected TB (Retroviral Tuberculosis). It proved that diagnosis extra pulmonary TB is more challenging

[26]. The most common case in Extra Pulmonary TB is lymphadenitis, i.e. inflammation of one or a few lymph nodes, which is around the neck, armpits and groin.

5. DATA MINING METHODS FOR TUBERCULOSIS DIAGNOSIS

Classification is one of data mining task that commonly used to analyze medical data [17]. Some researchers compared several methods in order to study which method obtained the highest accuracy in diagnosing tuberculosis. The study of [17] indicated that the rough neural networks technique gave better results compared to neural networks. They have chosen rough sets and neural networks because those methods can discover pattern in ambiguous and imperfect data and provide tools for data and pattern analysis. Rough set is also compared with ANFIS [18]. The result showed that ANFIS classified tuberculosis data with 97% of correctness, while rough set 92%. In addition, ANFIS is compared with Multilayer Perceptron and PART [24]. According to their study, ANFIS is an accurate and reliable method for classification of tuberculosis patients when compared with Multilayer Perceptron and PART algorithms.

Data mining approach can also be combined. For instance, clustering is combined with classification [22]. They used k-means to cluster TB data into two clusters, that is, Pulmonary TB and Retroviral TB. Then classified the instances using several algorithms such as k-NN, C.45 decision tree, Naive Bayes, RandomForest, Support Vector Machine, bagging and AdaBoost. The results showed that SVM obtained the highest accuracy, 98,7% compared to other classifiers.

Various data mining methods used by some researchers in diagnosing tuberculosis are summarized in Table 2.

Table 2. Various Data Mining Methods Used for Tuberculosis Diagnosis

Auth or(s)	Methods																	Result/Accu racy
	Lazy Modeling		Bayesian Modeling		Tree Classifier			Neural Network							SVM	Meta Modeling		
	k-NN	k- Means	NB	BN	C.45 DT	RF	PART	N N	R S	RNN	BP	Neuro fuzzy	ANFIS	MP		Bag- ging	Ada- Boost	
Bakar & Febriyani, 2007								√	√	√								RS 92,14% NN 90,44% RNN 92,29%
Benfu et al., 2009											√							Accuracy 93.10%, sensitivity 88.89% and specificity 100%
Asha et al., 2011	√	√	√		√	√									√	√	√	SVM has the highest accuracy 98,7%, followed by Bagging 98,4% and RandomFore st 98,3%
Uçar & Karahoca, 2011							√						√	√				RMSE: ANFIS 18%, MP 19% , PART 20%
Ali et al, 2012			√	√														Naive Bayes 90.7563% Bayesian Network 93,2773%
Ansar i et al., 2012											√	√	√					Training error using 70 training data: 0,0052404
Uçar et al., 2012									√				√					Correctness using ANFIS 97% while RS 92%

NB: Naive Bayes, BN: Bayesian Network, DT: Decision Tree, RF: Random Forest, NN: Neural Network, RS: Rough Set, RNN: Rough Neural Network, BP: Back Propagation, ANFIS: Adaptive Neuro Fuzzy Inference Systems, MP: Multilayer Perceptron

6. CONCLUSION



Literatures gathered from many sources such as IEEE explorer, Elsevier and ScienceDirect, are classified by variables, data preprocessing techniques and methods used for tuberculosis diagnosis. From those selected literatures that have been reviewed, we conclude that the most frequently used variables are sweating at night, more than 3 weeks of cough, fever, weight loss, age, and chest pain respectively. Support Vector Machine gave the highest accuracy 98,7%, followed by Bagging 98,4% and RandomForest 98,3% compared to other methods. In addition, several experiments have shown that ANFIS is the most accurate method in diagnosing tuberculosis.

REFERENCES

- [1] WHO, "Global Tuberculosis Report 2012," 2012.
- [2] X. Y. Djam and Y. H. Kimbi, "A Decision Support System for Tuberculosis Diagnosis," *The Pacific Journal of Science and Technology*, vol. 12, no. 2, pp. 410–425, 2011.
- [3] PDPI, "Tuberkulosis Pedoman Diagnosis dan Penatalaksanaan di Indonesia," 2002.
- [4] R. U. K. R. M. Radzi, W. Mansor, and J. Johari, "Review of mycobacterium tuberculosis detection," in *2011 IEEE Control and System Graduate Research Colloquium*, 2011, pp. 189–192.
- [5] M. K. Osman, M. Y. Mashor, and H. Jaafar, "Detection of Mycobacterium Tuberculosis in Ziehl Neelsen Stained Tissue Images using Zemike Moments and Hybrid Multilayered Perceptron Network," 2010, pp. 4049–4055.
- [6] R. Lestari, M. Ahmad, B. Alisjahbana, and T. Djatmiko, "The Lung Diseases Diagnosis Software : Influenza and Tuberculosis Case Studies in The Cloud Computing environment," in *International Conference on Cloud Computing and Social Networking*, 2012, no. January, pp. 1–7.
- [7] H. M. S. C. Kusuma, "Diagnostik Tuberkulosis Baru," *Sari Pediatri*, vol. 8, no. 4, pp. 143–151, 2007.
- [8] M. Bhatt, R. Bhaskar, and S. Kant, "Pulmonary tuberculosis as differential diagnosis of lung cancer," *South Asian Journal of Cancer*, vol. 1, no. 1, p. 36, 2012.
- [9] F. M. E. Uzoka, J. Osuji, and O. Obot, "Clinical decision support system (DSS) in the diagnosis of malaria: A case comparison of two soft computing methodologies," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1537–1553, Mar. 2011.
- [10] A. Kusiak, K. H. Kernstine, J. A. Kern, K. A. McLaughlin, and T. L. Tseng, "Data Mining : Medical and Engineering Case Studies," in *Proceeding of the Industrial Engineering Research 2000 Conference*, 2000, pp. 1–7.
- [11] F. M. E. Uzoka, J. Osuji, and F. O. Aladi, "A framework for cell phone based diagnosis and management of priority tropical diseases," *IST-Africa Conference*, pp. 1–13, 2011.
- [12] B. H. Tracey, G. Comina, S. Larson, M. Bravard, J. W. López, and R. H. Gilman, "Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis," in *33rd Annual International Conference of the OEEE EMBS*, 2011, no. day 0, pp. 6017–6020.
- [13] M. K. Osman, M. Y. Mashor, H. Jaafar, R. A. . Raof, and N. H. Harun, "Performance Comparison between RGB and HSI Linear Streching for Tuberculosis Bacilli Detection in Ziehl-Neelsen Tissue Slide Image," *2009 IEEE International Conference on Signal and Image Processing Applications*, pp. 357–362, 2009.
- [14] M. K. Osman, F. Ahmad, Z. Saad, M. Y. Mashor, and H. Jaafar, "A Genetic Algorithm-Neural Network Approach for Mycobacterium Tuberculosis Detection in Ziehl-Neelsen Stained Tissue Slide Images," in *2010 10th International Conference on Intelligent Systems Design and Applications*, 2010, pp. 1229–1234.
- [15] M. K. Osman, M. H. M. Noor, M. Y. Mashor, and H. Jaafar, "Compact single hidden layer feedforward network for mycobacterium tuberculosis detection," in *2011 IEEE International Conference on Control System, Computing and Engineering*, 2011, pp. 432–436.
- [16] H. Wang, C. Zhao, and F. Li, "Identification of M. tuberculosis complex by a novel hybridization signal amplification method," *Proceedings 2011 International Conference on Human Health and Biomedical Engineering*, pp. 1085–1088, Aug. 2011.
- [17] A. A. Bakar and F. Febriyani, "Rough Neural Network Model for Tuberculosis Patient Categorization," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, 2007, no. 1, pp. 765–768.
- [18] T. Uçar, A. Karahoca, and D. Karahoca, "Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets," *Neural Computing & Applications*, 2012.
- [19] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second. San Francisco, Canada: Morgan Kaufmann Publishers, 2006, p. 772.
- [20] H. Dağ, K. E. Sayın, I. Yenidoğan, S. Albayrak, and C. Acar, "Comparison of Feature Selection Algorithms for Medical Data," in *2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, 2012, pp. 1–5.
- [21] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Learning," *International Journal of Computer Science (IJCS)*, vol. 1, no. 2, pp. 111–117, 2006.

- [22] T. Asha, S. Natarajan, and K. N. B. Murthy, "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification," *Journal of computing*, vol. 3, no. 4, 2011.
- [23] Y. Benfu, S. Hongmei, S. Ye, L. Xiuhui, and Z. Bin, "Study on the Artificial Neural Network in the Diagnosis of Smear Negative Pulmonary Tuberculosis," in *2009 World Congress on Computer Science and Information Engineering*, 2009, pp. 584–588.
- [24] T. Uçar and A. Karahoca, "Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches," *Procedia Computer Science*, vol. 3, pp. 1404–1411, Jan. 2011.
- [25] T. Asha, S. Natarajan, and K. N. B. Murthy, "Data Mining Techniques in the Diagnosis of Tuberculosis," in *Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis*, P.-J. Cardona, Ed. InTech, 2012, pp. 333 – 352.
- [26] M. Bahadori and M. H. Azizi, "Common Challenges in Laboratory Diagnosis and Management of Tuberculosis," *Iranian Red Crescent Medical Journal*, vol. 14, no. 1, pp. 3–9, 2012.

BIBLIOGRAPHY OF AUTHORS

	<p>Rusdah, M.Kom</p> <p>She received her S1 degree (S.Kom) in Information System and M.Kom in Software Engineering from Universitas Budi Luhur. She is a Ph.D student at Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University (since 2012). She is a lecturer at Department of Information System, Faculty of Information Technology, Universitas Budi Luhur. Her research interest are decision support, data warehousing and data mining.</p>
	<p>Drs. Edi Winarko, M.Sc., Ph.D.</p> <p>He received his S1 degree in Statistics from Gadjah Mada University, MSc. in Computer Sciences from Queen's University, Canada, and Ph.D in Computer Sciences from Flinders University, Australia. He is a lecturer at the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. His research interest are data warehousing and data mining, and information retrieval as well.</p>