Advanced Science Letters
Vol. 23, 6527–6529, 2017

Support Vector Machine for Classification of Pulmonary Tuberculosis in Semarang

Moh. Yamin Darsyah[1] and Sri Darmawati[2]
[1]*Department of Statistics, University of Muhammadiyah Semarang, Jl. Kedungmundu Raya No.18. Semarang, 50273, Indonesia*
[2]*Department of Health Analyst, University of Muhammadiyah Semarang, Jl. Kedungmundu Raya No.18. Semarang, 50273, Indonesia*

Pulmonary tuberculosis (pulmonary TB) is an infectious disease caused by Mycobacterium Tuberculosis that attacks the lungs organs. Until recently pulmonary TB is still a large category of disease that cause the death in Indonesia, so that should get a special treatment in order to reduce the number of morbidity and mortality resulting. Data of health department showed Semarang as the most contributor to the city of pulmonary TB incidence rate in Central Java. Another contributing factor is environmental factors include physical environment, individual characteristics, and social environment. Trace it, it is necessary to do an analysis of the factors that affect the status of whether infected household members (ART) or not against pulmonary TB in Semarang is a descriptive Supports Vector Machine (SVM) Methode. Descriptive analysis aims to examine the characteristics of ART based on environmental factors, whereas SVM is to determine classification accuracy of pulmonary TB. The result accuracy of classification is 98%.

Keywords : SVM, Classification Accuracy, Pulmonary TB.

1. INTRODUCTION

Tuberculosis (TBC or TB) is an infectious disease caused by bacillus acid resistant or BTA with the full name of the Mycobacterium Tubercolosis Bacteria. This bacterium is a bacillus bacterium which is very strong so it takes a long time to treat. These bacteria often infect the lungs than other parts of the human body, so that cases of tuberculosis that often occurs in Indonesia is a case of pulmonary tuberculosis. [1,2]

Tuberculosis disease is usually transmitted through the air contaminated with the bacteria of Mycobacterium Tubercolosis that released when the patient coughs, whereas in children the source of infection is generally derived from adult tuberculosis. Tuberculosis bacillus is inhaled through the respiratory tract into the lungs, then bacill entered again into the lung lymphatics tract and spread to various organs of the body through the bloodstream. In addition to humans, animals can also be infected and transmited the tuberculosis disease to humans through its feces.[3]

Tuberculosis (TBC) is a health problem, both in terms of mortality, the number of incidence of disease (morbidity), as well as diagnosis and treatment or therapy. In 2014, the WHO Global Surveillance estimates that in Indonesia there are 583 .000 people with tuberculosis in each year w ith the amount of 262.000 BTA positive or incidence rate of approximately 130 per 100.000 of population and the mortality from tuberculosis are estimated to hit 140.000 population per year.[4] The result of Household Health Survey (SKRT) in 2014 showed that tuberculosis is the third cause of death after cardiovascular disease and respiratory disease in all age groups and number one of the infection classes. Especially tuberculosis cases occur in the productive working age, ie the age group of 15 to 49 years old who have an impact on the quality of human resources that could interfere with the family's economy, society, and the state.[3]

With the increasing cases of transmission of infectious pulmonary tuberculosis that has been reported at this time, hence need for a theoretical study on the determination of relevant variables that affect the incidence of pulmonary tuberculosis. It is intended that the amount of pulmonary tuberculosis patients in Indonesia can be minimized. Data from Department of Health in Semarang City 2010 showed that Semarang city as the largest contributor to the incidence of pulmonary tuberculosis (pulmonary TB) in Central Java. Therefore, this study is expected to determine the model that represents the variables that affect the incidence of pulmonary TB disease in the city of Semarang.

Previous research on variables that estimated to affect pulmonary tuberculosis include age factor, gender, level of education, occupation, smoking habits, room occupancy density, ventilation, housing conditions, air humidity, nutritional status, socioeconomic circumstances, and behavior.[5]

From these explanations it can be said that there are a lot of factors that affect the incidence of pulmonary TB so it is necessary to identify the factors that most influential to be used for prevention planning and treatment of pulmonary

1936-6612/2017/23/6527/003 doi.org/10.1166/asl.2017.9673

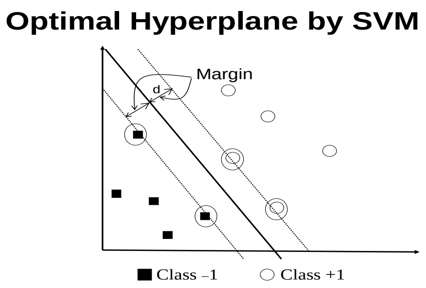
TB disease so that the incidence of this disease can be minimized. To describe the characteristics of household members (ART) which were infected with pulmonary TB and look at the most influential factors in a group then use the appropriate classification method based on the factors that influence it.[5]

Support Vector Machine (SVM) is a machine learning method that works on the principle of Structural Risk Minimization (SRM) with the goal of finding the best hyperplane that separates two classes in the input space.[6] In addition, SVM aims to minimize the upper limit of the general error. Another advantage of using SVM is that this method can be analyzed theoretically using the concept of computational learning theory.

2. EXPERIMENTAL DETAILS

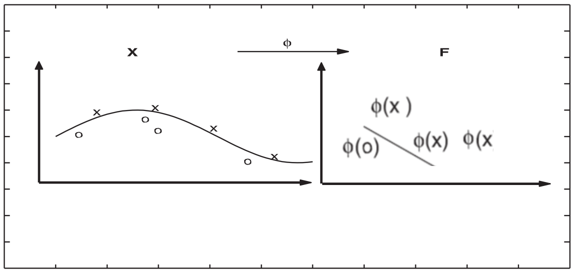
The basic principle of SVM is a linear classifier, then expanded to work on the nonlinear case by incorporating the concept of Kernels on high-dimensional workspace.[7] For example, given the equation of $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_m\}$ where the available data are denoted as $\mathbf{x}_i \in R^n, i = 1,2,...\text{m}$. Figure1 shows some of the patterns that are members of two classes, namely +1 and -1. Pattern belonging to class -1 is symbolized by a red box, while the pattern in class +1 is symbolized by a yellow circle. Classification problem can be translated to the effort of finding a line (hyperplane) that separates the two groups.

Figure 1. Optimal hyperplane by SVM



In general, in the real problem, the data of linearly separable are rarely found. So the Kernel function in Support Vector Machine is used to address the nonlinear data.[8,9] By entering the Kernel function, then the problem of nonlinear data become linear in the new space as shown in the following illustration.

Figure 2. Illustration of nonlinear problem becomes linear with Kernel SVM



Mathematically, some kernel functions are described as follows:

1. Linier Kernel: $x^T x$,
2. Polynomial Kernel: $\left(x^T x_i + 1\right)^p$,
3. Radial Basis Function (RBF) Kernel: $\exp\left(-\frac{1}{2\sigma^2}\|x - x_i\|^2\right)$,
4. Tangent Hyperbolic Kernel: $\tanh\left(\beta x^T x_i + \beta_1\right)$, where $\beta, \beta_1 \in \Re$

Logistic regression is a statistical method to learn about the pattern of mathematically relationship between one of the response variable (*y*) which is nominal or ordinal with one or more predictor variables (*x*). Response variable in logistic regression is binary or dichotomous variable. According to the type scale and the use of response variables, logistic regression was divided into 3 types, namely the binary logistic regression, multinomial, and ordinal.

Binary logistic regression analysis is a regression method that aims to determine the effect of the dependent variable (*y*) and the independent variable (*x*) where the *y* variable produces two categories, namely 0 and 1.[10] So that the *y* variable follows the Bernoulli distribution with probability function as follows:

$$f(y) = \pi^y (1-\pi)^{1-y} \;\; ; \;\; y = 0,1 \qquad (1)$$

Where if y = 0, then $f(y) = 1 - \pi$ and if y = 1, then $f(y) = \pi$. Logistic regression function can be written as follows:

$$f(z) = \frac{1}{1+e^{-z}} \; ekuivalen \; f(z) = \frac{e^z}{1+e^z} \qquad (2)$$

With $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

3. RESULT AND DISCUSSION

The data that used in this study are a secondary data obtained from the data of Basic Health Research (Riskesdas) and National Socioeconomic Survey (SUSENAS) 2014 in Indonesia were sourced from the Centre for Research and Development of Health Systems and Policy, Agency for Health Research and Development, Ministry of Health Republic of Indonesia.

The variables that used in this study consists of the dependent variable and the independent variables, as follows:

1. A member of the household (ART) is said infected to pulmonary tuberculosis if household member (ART) has ever tested positive for pulmonary tuberculosis infection in the span of the last one month before the survey and has been confirmed through blood tests by health professionals (doctors/ nurses/midwives).This is the dependent variable (Y) of binary scale with code 1 is provisions for household members (ART)-infected pulmonary tuberculosis and code 2 is for household members (ART) which is not infected with pulmonary tuberculosis.
2. Independent variables (X) that used in this study are all factors that influence the incidence of pulmonary tuberculosis in Semarang City, in terms of both the environment and people's behavior. The factors used in this study include:

Selected independent variables and estimated to explain the dependent variable are as follows:

a. Education (X_1); this variable indicates the highest educational status owned by ART. This variable has an ordinal scale with six categories, namely: code 1 if never attended school, code 2 if not finished elementary school, code 3 if finished elementary school, code 4 for junior high school education, code 5 for high school education, and code 6 for college.
b. Employment (X_2); hhis variable indicates the type of ART primary work. This variable has a nominal scale with eight categories: code 1 if not work, code 2 for workers (workers who earn wages in processing the work of others, such as farm workers, construction workers, laborers lift transport, and labor workers), code 3 for fishermen, code 4 for farmers, code 5 for the self-employed, code 6 for private employees, code 7 for civil servants (working in government as civil servants), and code 8 for the other code.
c. Socioeconomic Status (X_3); this variable has nominal scale which shows the economic status of ART with two categories, namely: code 1 indicate the poor status and code 2 for the nonpoor status.
d. Smoking Habit (X_4); this variable indicates the habit of ART in consuming tobacco products in the past one month (not only in the form of cigarettes, but also includes cigars, pipe, hand-rolled cigarettes, chewing tobacco). This variable has an ordinal scale with four categories: code 1 if smoking with daily frequency, code 2 if smoking with occasional frequency, code 3 if had never smoked but had previously smoked, and code 4 if never smoked at all.
e. Alcohol Consumption Habit (X_5); this variable indicates the habit of ART in consuming a drink containing alcohol in the past 12 months (branded alcoholic drinks, eg: beer, whiskey, vodka, wine, etc. as well as the traditional alcoholic drink, for example: *tuak*, *poteng*, *sopi*). This variable has a nominal scale with two categories: code 1 if consumed alcohol in the past 12 months, and code 2 if not consumed alcohol in the past 12 months.

Table I. Testing parameters simultaneously

	Chi-square	df	sig
Step	113.789	5	0.000
Block	113.789	5	0.000
Model	113.789	5	0.000

Based on Table I indicates that the testing parameters is simultaneously proving that all the independent variables affect the patients of pulmonary tuberculosis with very significant value (<5%).

Table II. Partial parameter testing

Variabel*x*	Sig.
1	0.056
2	0.043
3	0.051
4	0.039
5	0.043

Here's a partial test of variables that shown in Table II. Variables type of work (X_2), smoking habits (X_4), and alcohol consumption habits (X_5) are a variable that significantly affect patients with pulmonary tuberculosis while variable levels of education and socioeconomic status are not affect significantly (> 5%).

Table III. Partial Parameter Testing

Level of accuracy	SVM (RBF)	Logistic regression
Y	0.981	0.713

The level of accuracy in the classification of cases is very necessary because it indicates the precisely of a category of disease whether infected or not on the patient so as to determine the level of accuracy in the case of pulmonary tuberculosis needs to be done to detect the accuracy of the truth. In Table III shows the accuracy rate is so high at 98% in the classification method using SVM with RBF Kernel, if the accuracy precision in the classification is high so the medical treatment measures would be appropriate.

4. CONCLUSION

Variables type of work (X_2), smoking habits (X_4), and alcohol consumption habits (X_5) are a variable that significantly affect patients with pulmonary tuberculosis. The level of accuracy in the classification with SVM method obtained by 98%. For further research could use other classification methods so the level of accuracy is known for proper medical treatment.

Although the slow reduction of the incidence of TB has been seen in developed countries, TB is still a major challenge among infectious diseases, even in the 21st century. Fast and accurate TB testing, such as bacterial DNA analysis and whole-blood interferon-γ assay, has been developed for detecting latent infection. The traditional imaging concept of primary and reactivation TB has recently been challenged on the basis of DNA fingerprinting, and radiologic features depend on the level of host immunity rather than the elapsed time after the infection.

References and Notes

1. J.P Cegielski, D.P Chin, and M.A Espinal. The global tuberculosis situation: progress and problems in the 20th century, prospects for the 21st century. *Infect Dis Clin North Am*; 16:1–58 **(2002)**.
2. E.L Corbett, C.J Watt, and N. Walker. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* ; 163:1009–1021**(2003)**.
3. J.M Tufariello, J. Chan, and J.L Flynn. Latent tuberculosis: mechanisms of host and bacillus that contribute to persistent infection. *Lancet Infect Dis*; 3:578–590 **(2003)**.
4. World Health Organization. Fact sheet no. 104. Tuberculosis. www.who.int/mediacentre/factsheets/fs104. WHO Website. Revised March **(2014)**.
5. K.S Lee and J.G Im. CT in adults with tuberculosis of the chest: characteristic findings and role in management. *AJR*; 164:1361–1367 **(1995)**.
6. S. Abe. Support Vector Machines for Pattern Classification. London: Springer-Verlag **(2010)**.
7. T. Vance, N. Reljin, A. Lazarevic, D. Pokrajac, V. Kecman, N. Melikechi, A. Marcano, Y. Markushin and S. McDaniel. Classification of LIBS Protein Spectra Using Support Vector Machines and Adaptive Local Hyperplanes. IEEE. 1–7 **(2010)**.
8. C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. 2: 121–167 **(1998)**.
9. N. Bhardwaj, R. Langlois, G. Zhao, and H. Lu: Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res. 33(20):6486–6493 **(2005)**.\
10. C.T. Le. *Applied Categorical Data Analysis*. USA: John Wiley and Sons **(1998)**.