Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 130 (2018) 231–238

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Ambient Systems, Networks and Technologies, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, 8-11 May, 2018, Porto, Portugal

Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis

Rakhmetulayeva S.B.[a], Duisebekova K.S.[a,b],

Mamyrbekov A.M. [a], Kozhamzharova D.K.[a,*],

Astaubayeva G.N. [c], Stamkulova K. [c]

[a]*International Information Technology University, Almaty, Kazakhstan*
[b] *Al-Farabi Kazakh National University Almaty, Kazakhstan*
[c]*International Narxoz University, Almaty, Kazakhstan*

Abstract

In this paper the algorithm based on SVM (Support Vector Machines) for determining of effectiveness of test drug is proposed. The most time-consuming tasks of medicine include diagnosing and choosing a course of treatment. Traditionally, doctors have solved these problems, relying only on their own intuition and experience. Today in their arsenal are increasingly included methods based on high technology and allowing to process large flows of information. Mathematical description and formulation of the problem are given. The results of experiments based on the algorithm are presented. The main goal of the paper is to create prerequisites for preventive diagnosis of tuberculosis patients. Using the proposed model and monitoring system, specialists can correctly diagnose and develop an optimal treatment course.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: tuberculosis; support vertor machine; SVM; data analysis; onthology.

*Corresponding author. Dinara Kozhamzharova Tel.: +7-702-887-00-55; e-mail address: dinara887@gmail.com, d.kozhamzharova@iitu.kz

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2018.04.034

1. Introduction

Tuberculosis (TB) is an infectious disease caused by a bacillus called Mycobacterium tuberculosis. According to the report of World Health Organization's Global Tuberculosis Report, TB is the ninth leading cause of death worldwide and the leading cause from a single infectious agent, ranking above HIV/AIDS. In 2016, there were an estimated 1.3 million TB deaths among HIV-negative people (down from 1.7 million in 2000) and an additional 374 000 deaths among HIV-positive people. An estimated 10.4 million people fell ill with TB in 2016: 90% were adults, 65% were male, 10% were people living with HIV (74% in Africa) and 56% were in five countries: India, Indonesia, China, the Philippines and Pakistan.

Drug-resistant TB is a continuing threat. In 2016, there were 600 000 new cases with resistance to rifampicin (RR- TB), the most effective first-line drug, of which 490 000 had multidrug-resistant TB (MDR-TB).4 Almost half (47%) of these cases were in India, China and the Russian Federation. Globally, the TB mortality rate is falling at about 3% per year. TB incidence is falling at about 2% per year and 16% of TB cases die from the disease; by 2020, these figures need to improve to 4–5% per year and 10%, respectively, to reach the first (2020) milestones of the End TB Strategy. Most deaths from TB could be prevented with early diagnosis and appropriate treatment. Millions of people are diagnosed and successfully treated for TB each year, averting millions of deaths (53 million 2000–2016), but there are still large gaps in detection and treatment[1].

2. Diagnosis and treatment system

Tuberculosis is still a major problem in advanced countries due to specific socioeconomic factors. From a global perspective, many laboratories use the same methods today that were in use long time ago for the detection of tuberculosis, because most of innovative current technologies for the detection of tuberculosis incurs high cost and cannot be afforded for all the countries[2]. It can lead to death in untreated and inappropriately treated patients particularly in countries with low income. Therefore, early diagnosis of the disease not only increases treatment success, but also reduces death rates. Today, due to high classification and diagnosis rates, specialist systems have become an important tool in diagnosis of the disease. In this study, a support vector machine (SVM), which is a machine learning technique was used for preliminary diagnosis of tuberculosis disease for the first time[3]. Today, IT is used in medicine in three ways:

- use of equipment for surgical treatment, observation of the patient in preoperative and postoperative periods, etc.;
- maintenance of document circulation and financial accounting;
- predicting the state of the body, diagnosing diseases, monitoring the stages of the development of relapse, prescribing the necessary course of treatment with the help of intelligent decision-making systems.

Let's try to analyze the state of affairs in the third, the most knowledge-intensive direction. The systems of analysis and forecasting created in our country and abroad, artificial intelligence systems, designed for diagnosis, allow solving a variety of tasks. These tasks can be very narrow (for example, scheduling an effective intake of certain medications for a specific patient), and more global, related to predicting the patient's condition, issuing recommendations for surgery and analyzing the possible condition of the patient in the postoperative period. Depending on the purpose, such systems are based on different methods; among them:

- statistical methods of data processing (SMDP);
- artificial neural networks (ANN);
- nonlinear regression methods (NRM);
- reasoning on the basis of similar cases (RBSC).

Previously it was believed that the accuracy of the model can be improved solely by taking into account a number of factors and their composition. But this approach required more and more retrospective (i.e., the period of consideration of statistical data), which is often impossible to implement. The number of structural elements involved in the creation of the mathematical model was limited, which led to a statement of the existence of such a tabulated relationship (with which every medical researcher deals), which cannot be approximated with the composition of the selected set of elements. Medical research is usually faced with this situation: a large number of parameters with tabulated values are involved, and the relationship between them is clearly not visible. RBSC (Reasoning on the basis of similar cases) is a traditional option for decision-making by physicians, but the problem is how to store information on all cases that occurred in the practice of the doctor. It is necessary to systematize the data, select the parameters and their values to encode the information so that the required description can be extracted at any time. Moreover, this information, as a rule, does not have a mathematical or even a numerical expression that allows using the listed methods. Textual information, for example, a description of the reaction of the body to various stimuli or drugs, in general is not always possible to translate into quantitative. In such cases, you need to turn to the logic device, which provides manipulation of linguistic (text) variables. The presence of a variety of diagnostic methods significantly complicates their choice, on which the accuracy and completeness of the result depends. In the case of unreasonable choice, it may appear that all work has been unsuccessful, but with a comprehensive assessment of the type and structure of information, it is often possible to select a method that provides good results. So, if the information has a predominantly numerical expression, then the best results will be shown exactly by those methods in which the mathematical expression of the studied process is used. If the data is descriptive, the best way is to perform logical processing. Using the system, the physician can get a more complete picture of the disease, which will help him diagnose. Based on the information from the predictive system, further steps can be outlined, for example, to determine which additional tests or examinations need to be conducted in order to obtain new arguments in favour of a particular diagnosis. If the information base is correctly constructed (the optimal number of auxiliary fields providing the maximum speed of analysis and screening of unsuitable cases), you can view the complete medical history of even unusual cases that occurred in the practice of a doctor or clinic.

The system of forecasting described in the article not only allows more accurate diagnosis of diseases, but also helps in choosing a course of treatment. Now there is work to improve the accuracy of determining the course of treatment.

The system based on the study the purpose of which is to prove the effectiveness of treatment using test drug by experimenting in humans. The experiment lasted for 3 years amount 585 patients, and the average duration of the experiment per 1 patient was 6-12 months. For each patient was conducted individual records visits and medication.

3. Used Methods

It is supposed to use linear classification algorithms, in particular SVM. There is given a training sample

$$x^l = (x_i, x_i)_i^l = 1, l = 66 \tag{1}$$

where $x_i \in R^n, n = 20$,

$$y_i \in \{A_1, A_2, B_1, B_2\} \tag{2}$$

For each of the tasks of the two-class classification (separating one class from the other three and separating pairs of classes from each other), we will re-encode the classes so that $y_i \in \{-1,1\}$. It is required to select the parameter vector of the optimal separating hyperplane ω, which minimizes the functions of the sliding control:

$$LOO(\omega, X^l) = \sum_{i=1}^{l} \left\lfloor a(x_i, X^l \setminus x_i, \omega) \neq y_i \right\rfloor \longrightarrow \min_{\omega} \quad (3)$$

where $a(x) = \left\lfloor \sum_{j=1}^{n} \omega_j x^j - \omega_0 > 0 \right\rfloor$.

A.Basic Assumptions

A particular feature of this task is the large dimensionality of the feature space and a small number of precedents. Thus, in order to avoid retraining and to achieve a stable classification, it is required to solve the task of selecting the characteristics. For this purpose, the Relevance Kernel Machine with supervised selectivity algorithm (hereinafter - $\mu - RKM$) is used, which combines the possibilities of solving the classification and features selection problem.

B.Mathematical description of algorithms

a) Quasi-probabilistic formulation of the problem

Let Ω be the set of objects, each of which belongs to one of two classes: $(\omega) \in Y = \{-1,1\}$ (ω). Each object $\omega \in \Omega$ is characterized by $'n'$ signs in some scales $x^2(\omega) \in X_i$. Suppose that some unknown hyper plane is objectively defined in the space of features=X_1*…*X_n . As a model of object distribution, we will consider two improper parametric distributions:

Next, we will consider a vector $(V_1,\ldots,V_n,b)$ as a random vector with an a priori distribution density $\Psi(V_1,\ldots,V_n,b)$ By the Bayes formula, the posterior density of the distribution of the parameters ϑ and b .

Selection $X_1,\ldots,X_n$ is obtained from the general population with the distribution function Ψ (x | υ). Let ψ (x | υ) be the distribution density of the observed random variable ξ, if ξ is continuous or probability P (ξ = X | u), if ξ is discrete, provided that the value of the unknown parameter is υ. The likelihood function L (X_1, ..., X_n | u) of the available data is determined by the relation L (X_1, ..., X_n | u) = ψ (X1 | υ) ψ (X2 | u). . . · Ψ (Xn | υ). The calculation of the a posteriori distribution p~ (υ | X_1, ..., X_n) is carried out using the Bayesian formula (4), where Ai is the event that the value of the parameter being evaluated is υ, B is the event consisting in the fact that the values of n observations are fixed at the levels $X_1,\ldots,X_n$.

b) Method $\mu - RKM$ [4]

The Bayesian approach is based on two propositions. The degree of our confidence in the validity of some statement is numerically expressed in probability. When making a decision, two types of information are used as the initial information: a priori and contained in the initial statistical data. The a priori information is presented in the form of some a priori probability distribution of the unknown parameter being analyzed, which describes the degree of its confidence that this parameter will take some value, even before the collection of the initial statistical data. As the initial statistical data arrive, this distribution is refined, moving from a priori distribution to a posteriori distribution, according to the Bayes formula (4).

Let the a priori distribution densities of the components of the directing vector of the separating hyper plane ϑ_i have normal distributions with zero mathematical expectations and variances

$$r_i : \varphi(\vartheta_i \mid r_i) = \frac{1}{\sqrt{2\pi r_i}} \exp\left(-\frac{1}{2r_i} K_i(\vartheta_i, \vartheta_i)\right) \quad (4)$$

We will assume that the parameter b has a uniform improper distribution, equal to a unity on the whole numerical axis. Then, the density of the vector distribution ϑ is proportional:

Suppose that all quantities $\frac{1}{r_i}$ have an a priori gamma distribution:

$$\gamma\left(\frac{1}{r_1} \mid \alpha, \beta\right) \infty \left(\frac{1}{r_i}\right)^{\alpha-1} \exp\left(-\beta\left(\frac{1}{r_i}\right)\right) \quad (5)$$

Assume that $\alpha = \frac{(1+\mu)^2}{2\mu}, \beta = \frac{1}{2\mu},$ where μ is some nonnegative parameter.

The principle of maximizing joint a posteriori density leads to the learning criterion:

For each iteration with a fixed approximation $\left(r_i^k, i = 1,\ldots,n\right)$ the solution of this optimization problem is reduced to only a small modification of the classical SVM.

If the current approximation $\left(V_i^k,\ldots,V_n^k, b^k\right)$ is found, then the next approximation $\left(r_i^{k+1},\ldots,r_{n1}^{k+1}\right)$ can be found from a simple relation:

$$\left(r_i^{k+1} = \frac{K_i(\vartheta_i,\vartheta_i) + \frac{1}{\mu}}{\frac{1}{\mu} + 1 + \mu} \right) \tag{6}$$

c) Characteristics selection

To increase the stability of the algorithm, an additional selection of characteristics was applied. For this, the RKM algorithm was preliminarily started and the received weight vector of the signs was considered:

$$\left(r\_i\right)\hat{}, i = 1,\ldots,n \tag{7}$$

Further, from the set of characteristics, those of them for which

$$\frac{\max_j \hat{r}_j}{\hat{r}^i} > \gamma, i = 1,\ldots,n \tag{8}$$

where γ is an additional parameter of selectivity, were deleted. On the modified feature set, the RKM algorithm was run again to obtain the final classification.

Variants or modifications

Parameters of the algorithm are $\in[0,\infty], \mu \in[0,\infty], \gamma \in[1,\infty]$, the best values of which are required to be selected as a result of the experiment.

4. Experiment

The initial data was collected in the database (See Figure 1), which consists of 1261 lines of data on health status in different periods, taking into account the number of visits. Each row of the table represents a one-time examination of the patient. During the survey, the following parameters were measured:

- identification parameters (ID) of the patient;
- number of visits to the patient;
- results of a patient's examination;
- group to which the patient belongs.

ID	Visit Name	result	f_Group
xx-001	Visit 16	success	placebo
xx-001	Visit 17	success	placebo
xx-001	Visit 15	success	placebo
xx-001	Visit 14	success	placebo
xx-001	Visit 13	success	placebo
xx-001	Visit 9	success	placebo
xx-001	Visit 8	success	placebo
xx-001	Visit 7	success	placebo
xx-001	Visit 6	success	placebo
xx-001	Visit 4	success	placebo
xx-001	Visit 3	success	placebo
xx-001	Visit 2	success	placebo
xx-001	Visit 12	fail	placebo
xx-001	Visit 11	success	placebo
xx-001	Visit 10	success	placebo
xx-001	Visit 1	success	placebo
xx-001	Screening	fail	placebo
xx-001	Visit 5	success	placebo
xx-003	Screening	fail	test_drug
xx-005	Visit 1	success	test_drug
xx-005	Visit 5	success	test_drug

Fig. 1. Structure of initial base segment

To construct tables using the data in database, we used Gretl software. This software also gives users an ability to write his or her own functions, which greatly expands the usefulness of the application.

This task was solved in several steps:

- data cleaning;
- transformation of initial data for further analysis;
- constructing tables in Gretl format for converted data;
- construction for the transformed data of histograms, graphs of means and other graphs in Gretl format.

We used our mathematical model and the data without results in order to train the system so that it would give the same results as real.

In all tables, the data were obtained using regression analysis in the Gretl open-source statistical package. A prediction model has been developed and the data grouped according to the patient's ID and divided into several clusters:

1) patients receiving advice and medication constantly (without interruptions);

2) patients rarely visiting a doctor and rarely taking a medicine.

	Coefficient	St.err	z	P-value	
VisitName	0, 131019	0,0151570	8,644	<0,0001	***
f_Group	-0,553546	0, 117679	-4,704	<0,0001	***
cut 1	0,801611	0, 251476	3, 188	0,0014	***

Average dependent change	1,503775		Standard deviation	0,500191
Logical likehood	-824,8586		The Akaike criterion	1655,717
Schwarz'Criteroin	1671, 136		The Hannan-Quinn(YQ)	166,511

Fig. 2. Structure of initial base segment ordered logit, used observations – 1261, Dependent variable: result, Standard error – QML

The number of 'correctly predicted' cases= 778 (61,7%)

Criterion for the likelihood ratio: Chi-square[5]= 98,3666 [0,0000]

Based on the obtained real data on patients (Figure 2), the estimated model is constructed (Figure 3). Further, the adequacy of the evaluated model was investigated for use in further forecasting. By obtained result we can assert with 95% accuracy that this model can be used to forecast in the long term (for example to determine the treatment schedule for a specific patient).

Observ	Result	Forecast	St.err	95% CI
1:1	1,00000	1,21069	0,483153	(0,262818, 2,15857)
1:2	1,00000	1,24162	0,482894	(0,294251, 2,18898)
1:3	1,00000	1,27254	0,482658	(0,325637, 2,21945)
1:4	1,00000	1,30347	0,482447	(0,356977, 2,24995)
1:5	1,00000	1,33439	0,482259	(0,388269, 2,28051)
1:6	1,00000	1,36532	0,482096	(0,419515, 2,31112)
1:7	1,00000	1,39624	0,481956	(0,450713, 2,34177)
2:1	1,00000	1,42716	0,481841	(0,481865, 2,37246)
2:2	1,00000	1,45809	0,481749	(0,512969, 2,40321)
2:3	1,00000	1,48901	0,481682	(0,544026, 2,43400)
2:4	1,00000	1,51994	0,481638	(0,575036, 2,46484)
...				
180:4	2,00000	1,57629	0,481910	(0,630857, 2,52173)
180:5	2,00000	1,70549	0,481882	(0,760105, 2,65087)
180:6	2,00000	1,67456	0,481782	(0,729379, 2,61975)
180:7	2,00000	1,67456	0,481782	(0,729379, 2,61975)
181:1	2,00000	1,70549	0,481882	(0,760105, 2,65087)
181:2	2,00000	1,57629	0,481910	(0,630857, 2,52173)
181:3	2,00000	1,70549	0,481882	(0,760105, 2,65087)
181:4	2,00000	1,70549	0,481882	(0,760105, 2,65087)
181:5	undefined	undefined	undefined	
181:6	2,00000	1,70549	0,481882	(0,760105, 2,65087)
181:7	2,00000	1,70549	0,481882	(0,760105, 2,65087)

Fig. 3. For 95% confidence intervals, t (1258, 0.025) = 1.962

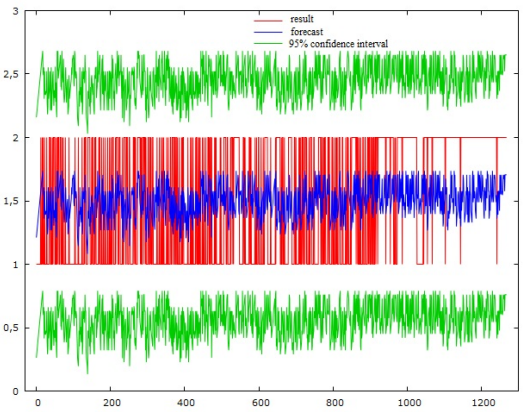

Fig. 4. Visualization of experimental data

In Fig. 4, visualization of the experimental data is presented, in which one can see the correctness of the constructed model, where there are no sharp jumps and stability and stability of the constructed model is observed.

5. Discussion

The system of forecasting described in the article allowed to create prerequisites for preventive diagnosis of tuberculosis patients, however the obtained results based on neural networks, developed according to the model described in the article does not satisfy us, in the future it is planned to improve the results by experiment with a large number of samples. Now there is work to improve the accuracy of determining the course of treatment, unfortunately, the facts of using similar systems in Kazakhstan are unknown.

6. Conclusion

Drug testing is an important application of statistical methods. Static analysis is an indispensable tool that allows practicing physicians and scientists to organize a study, calculate the sample size, assess the dose/effect relationship, visualize the data, build a variety of graphs and charts, and confirm the results. Thus, with the help of new information technology, the doctor takes a step forward towards evidence-based medicine. The proposed model helps to predict a possible improvement time for the patient, but because the data is poorly structured, it was decided to use a neural network. Using the previous successful experience of implementing models to process the Big Data[7,8], it is planned to obtain data using more samples on learning based on neural network.

References

1. "WHO | Global tuberculosis report 2017". http://www.who.int/tb/publications/global_report/gtbr2017_main_text.pdf
2. Pandiyan, O. El-Hassan, A. Hassan Khamis, and P. Rajasekaran, "Ontology with SVM based diagnosis of tuberculosis and statistical analysis". International Journal of Medical and Health Sciences Research, 2016,Vol.3, No.3, pp.37-43, ISSN(e): 2313-2752, ISSN(p): 2313-7746,DOI: 10.18488/journal.9/2016.3.3/9.3.37.43, 2016 Asian Medical Journals. All Rights Reserved.
3. A. Yahiaoui, O. Er, and N. Yumusak, "A new method of automatic recognition for tuberculosis disease diagnosis using support vector machines," Biomedical Research, Volume 28, Issue 9, 2017

4. S.Friedhelm, R.Fabio, K.Josef, "Multiple Classifier Systems," Proceedings of 12th International Workshop, MCS 2015, Günzburg, Germany, June 29 - July 1, 2015.
5. I. Campbell, "Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations," Statistics in Medicine, 26, pp. 3661–3675, (2007). DOI: 10.1002/sim.2832
6. R.L. Plackett, "Karl Pearson and the Chi-squared test," International Statistical Review, Longman Group Limited/Printed in Great Britain, International Statistical Institute, 51 (1983), pp. 59-72.
7. K. Duysebekovaa, V. Serbin, A. Kuandykov, T. Duysebekov, M. Alimanova, S. Orazbekov, D. Kozhamzharova, L. Alimzhanova, "The Solution of Semi-empirical Equation of Turbulent Diffusion in Problems of Polluting Impurity Transfer by Gauss Approach," The 11th International Conference on Future Networks and Communications (FNC) / 13th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), Montreal, Canada, August 15-18, 2016, Book Series: Procedia Computer Science, Volume: 94, pp. 372-379, 2016.
8. K. Duysebekovaa, V. Serbin, G. Ukubasova, Zh. Kebekpayeva, S. Aigul, S. Rakhmetulayeva, A. Shaikhanova, T. Duisebekov, D.Kozhamzharova, "Design and development of automation system of business processes in educational activity," Journal of Engineering and Applied Sciences, (8): pp. 4702-4714, 2017, ISSN:86-949X, Medwell Journals.