Classification of medicine characteristic using Super Vector Machine (SVM) at Palopo regional public Hospital Sawerigading

N Nirsal¹, S Paembonan², F N Yasir¹ and V B Dusmin¹

¹Universitas Cokroaminoto Palopo, Palopo (Indonesia)
²Universitas Andi Djemma Palopo, Palopo (Indonesia)

E-mail: nirsal@uncp.ac.id, solmin.p@gmail.com, fajarnovriansyah@uncp.ac.id, vickydusmin@uncp.ac.id

Abstract. One of the implementation of machine learning in medical world is to analyze medical dataset. Medical dataset used in this research was by using medicine dataset. Support Vector Machine Method is classification method of supervised learning, its algorithm works by using nonlinear mapping to change the data of real training into higher dimension. Selection of SVM Method is as solution to classify the characteristics of medicine. This method has function to make some similar medicines to be a group of certain data. The aim of this research was to obtain classification model which has high accuracy or small error in conducting classification of medicine data. Based on testing conducted, medicine classification using SVM produced accuracy=0,87.

1. Introduction

Medicine is defined as substance used in preventing and healing disease, it is also for healing and improving health for its user. Medicine has a very important role in health treatment. Some of medicine choices are available nowadays, so we need accurate considerations in selecting medicine for a disease. So many kinds of medicine available also can give apart problem in practice, especially about how to select and use medicine correctly and safely.

In conducting classification of medicine, one of the functions of information is as consideration matter in taking a decision or certain policy. There are some ways to produce useful information from data collection saved, it is by obtaining knowledge (information) saved in the data collection which is called data mining. From the explanation it can be analyzed about what characteristic develop every data in the data collection. It is needed a system which can give information in classifying medicine. The method used is supervised machine learning algorithm called SVM.

Machine learning is computer application focuses on development using artificial intelligence. SVM is one kind of machine learning method that are good to dealing with multi dimensions and continuous features. In this research, it is used SVM classification method to divide characteristics of many medicines into some groups so it makes easier to describe the characteristics of each of medicine group. Features of medicine used are composition, kind, and unit. Kotsiantis showed in his research that the accuracy of SVM method is bigger than other classification method, it has quickness of high classification, and high tolerance to irrelevant attribute [1]. The research conducted by (Isah et al,
2014) has conducted research related to finishing classification problem of user complain about the use of medicine and cosmetic by using SVM method [2]. Then, (Srivastava, 2011) in her research told that SVM had high accuracy to classify the pattern of network data packet [3].

2. Description of System

Data collected and prepared are data which are not arranged yet, not contain useful new information, so it is needed to do some processes to obtain useful information from the data. It needs a group of data which are so many so it is easier to obtain the expected pattern [4]. Algorithm of SVM presents document as vector and searches limitation called decision hyperplane, which separates two vector groups. Margin is used to measure the quality of decision hyperplane determined by space of two hyperplanes, which is parallel with decision hyperplane and touch the closest object from every vector namely support vector. The best decision hyperplane has maximum margin and related to all objects separated from each vector. In this research, it used SVM method, where this method produces good performance accuracy. In SVM, it tries to find out the best hyperplane of input space. This hyperplane is partition between two classes of input space [5]. The last result of SVM method is, obtaining medicine classification. The selection of this method is used to determined medicine classifications, which are based on composition, kind and unit of medicine. This research discusses about medicine classification of medicine data obtained from regional public hospital of Sawerigading. The procedures conducted in this research are:

3. Text Mining

Text mining adopts and develops technique from another aspect in accomplishing problem. Standard techniques usually used in text mining are text classification process, text clustering, ontology and taxonomy creation, document summarization, and latent corpus analysis [6].

Text mining and text analytics are coverage from wide data mining, then it is united in the same need, it is to change texts into numeral so the technique can be applied into big document database [7]. Generally, text mining must pass some processes to get expected result. The processes that must be conducted in text mining are [8]:

3.1. Text preprocessing.
This text beginning process aims to prepare text before it is processed further. The aim of text preprocessing process is to change raw input data into format which is suitable with next analysis.

3.2. Text transformation.
The process of text transformation conducts word change into basic word (stemming), decrease of word dimension of document (stop word removal), bag of word, and vectorial document representation for document representation.

3.3. Feature selection.
In this step, it is conducted discarding parts which are not related to get important parts.

3.4. Pattern discovery.
This step aim to find out pattern or knowledge from text by using data mining techniques like classification and clustering.

3.5. Interpretation.
This step conducted interpretation into forms to do evaluation.
4. Classification

In this research, it is used classification method to make data classification. The method will learn sample data using classification algorithm by knowing certain pattern of sample data to target class, so it enables to predict target class by using data out of sample data by using model classifier [4]. In classification, there are two steps, in the first step the model classifier will be formed based on data set or data training using classification algorithm, this process is called learning step. Then, model classifier is used to determine class label based on attribute [4].

Data training can be presented by using database, where every attribute is placed on each column in database. The second step is conducted evaluation of model classifier to obtain accuracy score. If accuracy score is suitable with what is expected, so this classifier can be used to get the result of prediction by using data out of training data [4]. The chart of classification process of sample data using model classifier to get prediction result [4] can be seen on Figure 1.

5. Support Vector Machine (SVM)

Support Vector Machine Classifier is a method used in developing a hyperplane or set hyperplane in high room or unlimited dimension, which can be used for classification, regression, or other jobs. Intuitively, a good separation can be achieved by hyperplane which has the biggest distance to the spot of the nearest training data from every class (called functional margin), because generally more bigger the margin, the lower general error from classifier.

Figure 2 presents some alternatives of separator area which can separate all datasets based on their class. Nevertheless, the best separator area does not only separate data but also it has the biggest margin [9].
Figure 2. Alternative of separator area (left) and the best separator area with the biggest margin (m) (right)

The data in this margin area is called support vector. Two classes can be separated by a pair of margin area which is parallel. The first margin area limits the first class while the second margin area limits the second class, so it is obtained:

\[ x_i \cdot w + b \geq +1 \text{ for } y_i = +1 \] (1)

\[ x_i \cdot w + b \leq -1 \text{ for } y_i = -1 \] (2)

\( w \) is normal area and \( b \) is position of relative area to coordinate center.

Margin score (distance) between margin area (based on formula of line distance to center spot) uses formula as follow:

\[ \frac{1 - b - (1 - b)}{w} = \frac{2}{|w|} \] (3)

Search of the best separator area with the biggest margin score is formulated to problem of constrain optimization with:

\[ \min \frac{1}{2} |w|^2 \] (4)

The problem of constrain optimization can be changed to:

\[ \min_{w,b} L_p (w, b, a) = \frac{1}{2} |w|^2 - \sum_{i=1}^{n} a_i y_i (x_i \cdot w + b +) + \sum_{i=1}^{n} a_i \] (5)

By minimizing \( L_p \) to \( w \) and \( b \), so from \( \frac{\partial}{\partial b} L_p (w, b, a) = 0 \), it is gotten:

\[ \sum_{i=1}^{n} a_i y_i = 0 \] (6)

and, from \( \frac{\partial}{\partial w} L_p (w, b, a) = 0 \), it is gotten:

\[ w = \sum_{i=1}^{n} a_i y_i x_i \] (7)

Substitution of formula 7 to \( L_p \) is obtained two problems LD with different constrain, using formula as follow:

\[ L_p (a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} a_i a_j y_i y_j x_i x_j \] (8)
Search of the best separator area can be formulated as follow:
\[
\max_a L_p = \sum_{i=1}^n a_i y_i (x_i \cdot w + b +) + \sum_{i=1}^n a_i
\]
subject to \( \sum_{i=1}^n a_i y_i = 0, a_i \geq 0 \) (9)

There is score of \( \alpha_i \) for every training data. Training data which has score \( \alpha_i > 0 \) is support vector while the rest has score \( \alpha_i = 0 \). So the function of decision produced is only affected by support vector. The class of \( x \) testing data can be determined based on score from function of decision with formula:
\[
f(x_d) = \sum_{i \in \text{support}} a_i y_i x_i x_d + b
\]
(10)
x \( i \) is support vector, ns = number of support vector and \( x_d \) is data that will be classified. With dot product \( x_i x_d \) it is often substituted with symbol \( K \), \( K \) is matric kernel, using formula:
\[
f(x_d) = \sum_{i \in \text{support}} a_i y_i k(x_i x_d) + b
\]
(11)
The function of separator used formula as follow:
\[
f(x_d) = \text{sign} \left( \sum_{i \in \text{support}} a_i y_i k(x_i x_d) \right) + b
\]
(12)

6. System Architecture

The procedure of process conducted in this research is described in flowchart form, which can be seen on Figure 3. The steps in medicine classification used SVM as follow:

1. Input all medicine data
2. Pre-processing, transformation, feature selection, interpretation, and pattern discovery are conducted for data of medicine text
3. Make manual label to medicine data
4. Classify medicine data using SVM method
5. Show the result of medicine classification
7. Results

7.1. Discussion of clustering result
The result of classification is observed by using SVM. Dataset used in this research is obtained from Regional General Hospital Sawerigading. The result of classification visualization can be shown in the form of scatter graphic visualization. Drug data totaling 890 datasets. The types of drugs shown in Figure 4 have labels including cream, gel, infusion, injection, caplet, capsule, lotion, shampoo, syrup, and tablets. From the classification results in Figure 4, it can be seen that the data on the type of tablet medicine has the most dominant membership classification. The classification result on figure 4 presents information of composition and kind data.
Information of kind and unit data classification can be seen on figure 5. The drug unit consists of the label gr, mcg, mg, and ml. The information displayed is the group of cream types in gr and mg units, gel and gr types, infusion type data in mg and ml units, injection in gr, mg and ml units, capsules with mg, then syrup type with mg and ml units, units of mcg and mg in combination with the type of tablet.

![Figure 5. The result of data classification of kind and unit of medicine](image)

Furthermore, in Figure 6, displays the classification results from the composition and unit data. The group with mg units with the composition of the drug has the most membership classification. Figure 6 shows the classification for drug composition in units of ml, gr, and mcg.

![Figure 6. The result of composition data classification](image)

7.2. Test of classification result
Test classification used accuracy method. Calculation of accuracy result of medicine classification is as follow:
\[
\text{accuracy} = \frac{65}{74} \times 100\% = 0.87
\]

8. Conclusion
Judging from the results of the research which resulted in an accuracy score of 0.87, it can be concluded that the SVM method can be used as a method for carrying out the drug classification process. The ability of the SVM method to divide into several categories will help accelerate the process of grouping drugs at the Sawerigading Palopo Hospital.
Acknowledgments
The author's gratitude goes to the Rector of Cokroaminoto Palopo University, Vice Rector for Academic Affairs, Director of Research and Community Services of Cokroaminoto Palopo University for moral and administrative support in writing this article, Dean of the Faculty of Computer Engineering who has directed the author in completing this article, Head of the Department of Informatics at Cokroaminoto Palopo University, Director of Sawerigading Palopo Hospital, and all parties who have contributed thoughts in this research activity.

9. References

[1] Kotsiantis S B 2007 Supervised Machine Learning: A Review of Classification Techniques (Informatica 13) p 268
[2] Isah H Trundle P dan Neagu D 2014 Social Media Analysis for Product Safety Using Text Mining and Sentiment Analysis (Computational Intelligence (UKCI) 2014 14th UK Workshop Vols. 1-7).
[3] Shrivastava K S dan Jain P 2011 Effective Anomaly Based Intrusion Detection (International Journal of Computer Vol. 3) p35-41.
[4] Han J and Kamber M 2006 Data Mining Concept and Technique (San Fransisco : Morgan Kaufman Publisher)
[5] Vapnik V N 1999 An Overview of Statistical Learning Theory (IEEE Transaction on Neural Networks Vol. 10) p5.
[6] Feinerer I Hornik K dan Meyer D 2008 Text Mining Infrastructure in R (Journal of Statistical Software Vol. 25) p1-54.
[7] Miner G Delen D Elder J Fast A Hill T dan Nisbet R 2012 Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications (Elsevier Inc).
[8] Even Y dan Zohar 2002 Introduction to Text Mining, Automated Learning Group National Center for Supercomputing Applications (University of Illionis)
[9] Sembiring K 2007 Penerapan Teknik Support Vector Machine untuk Pendeteksian Instrusi pada Jaringan (Skripsi, Program Studi Teknik Informatika ITB)