JITECS 113

by 113 Jitecs

Submission date: 21-May-2019 12:03PM (UTC+0700)

Submission ID: 1133725268

File name: 113-484-1-SM.docx (122.94K)

Word count: 2877

Character count: 14363

Journal of Information Technology and Computer Science
Volume 1, Number 1, 2016, pp. xx xx
Journal Homepage: www.jitecs.ub.ac.id

Classification Tuberculosis DNA using LDA-SVM

Received xx month xxxx; accepted xx month xxx

Abstract. Tuberculosis is a disease caused by the mycobacterium tuberculosis virus. Tuberculosis is very dangerous and it is included in the top 10 causes of the death in the world. In its detection, errors often occur because it is similar to other diffuse lungs. The challenge is how to better detect using DNA sequence data from mycobacterium tuberculosis. Therefore, preprocessing data is necessary. Preprocessing method is used for feature extraction, it is k-Mer which is then processed again with TF-IDF. The use of dimensional reduction is needed because the data is very large. The used method is LDA. The overall result of this study is the best *k* value is k = 4 based on the experiment. With performance evaluation accuracy = 0.927, precision = 0.930, recall = 0.927, F score = 0.924, and MCC = 0.875 which obtained from extraction using TF-IDF and dimension reduction using LDA.

1 Introduction

Tuberculosis (TB) is a dangerous disease. Based on WHO (World Health Organization), TB is one of the top 10 causes of death in the world. TB ever occupied the second cause of death after HIV/AIDS. This means that TB has emerged as a global health threat in this century [1]. TB is caused by the mycobacterium tuberculosis virus. In addition, the TB virus has resistance to drugs and not all TB can be treated with the same drug. Like tuberculosis with lineage from Beijing which has the highest resistance to drugs so it requires different treatments [2]. Drug-resistant TB (DR-TB) is a major threat because in 2013 as many as 3.5% of patients with tuberculosis had a defense against the drugs given [3]. The current challenge is how to detect and treat this TB. In some case the disease is difficult to detect because it resembles other respiratory diseases [4]. Therefore, it is necessary to detect TB better by using DNA data from mycobacterium tuberculosis because each organism must have DNA that differentiates and characterizes the organism. In its detection, it can be implemented by machine learning algorithms and included in the branch of science of bioinformatics. The use of ML is very important for bioinformatics because it can learn and build predictive models from input in the form of genomes, proteins, etc. then analyze it [5].

There are various variants of the mycobacterium tuberculosis virus and each DNA has a different length of data. Mycobacterium tuberculosis DNA contains a sequence of nitrogen base codes {A, T, C, G} with that order reaching thousands [6]. important aspects before classification are the selection features from DNA sequence. The research focuses on how to extract features from DNA sequence data and preprocess it before classification. After that, dimension reduction is done to reduce large data dimensions so that the classification process will be faster. The main focus of this research is how to preprocess data with the right method.

For feature extraction from DNA sequence data using k-Mer. To get uniform data length using TF-IDF. The data used is a complete genome that has up to thousands of long data, it should be high dimension of data. Therefore, dimension reduction will be carried out using LDA. The ML algorithm used is SVM because it is able to classify well.

The use of k-Mer has been successfully applied to similar studies that use sequence-based data such as DNA [7]. With k-Mer can provide stable and sometimes low accuracy depending on the selection of the appropriate k value because each value k contain different information [8]. This study we used TF-IDF to change text data to numerical. TF-IDF can convert data from DNA substring to matrix. TF-IDF converts data based on the frequency of occurrence of words in one. In this study we used LDA to reduce the data dimension. These techniques applied for feature extraction from large dimension data to lower dimensions. LDA process is based on supervised learning [9]. In previous research have used these methods but with different objects. So, the output of the research is to prove and get the right k value from k-Mer to extract the feature. Machine learning method that will be used in this study is SVM because it has been successfully applied to many classification problems. The advantages of SVM can avoid overfitting and being able to generalize data properly. So, in this research uses SVM as its classification method [11] [12].

2 Method

2.1 K-Mer

```
TTGACCGATGACCCT
TTGACC
 TGACCG
  GACCGA
   ACCGAT
    CCGATG
     CGATGA
      GATGAC
       ATGACC
        TGACCC
         GACCCT
```

Figure 1. Sample of k-Mer

In terms of biological sequences, k-mer can be defined as all possible subsequences with length of k [12]. In other words, the substring generated from k-mer can represent the entire length of the data sequence as shown in Fig.1. In the field of bioinformatics, k-Mer is used as feature extraction especially for metagenome analysis. Extraction of features from k-Mer is based on the frequency of the occurrence of the combination forming the DNA.

How k-Mer works is quite simple, k-Mer will take the substring based on the specified k value. Increasing the value of k will produce data with large dimensions and requires high time. The k value of k-Mer greatly affects the features produced. In this study we will try feature extraction with different k values, starting from 2 to 9.

2.2 Term Frequency – Inverse Document Frequency (TF-IDF)

After extracting features using k-Mer, then changing the data into numbers and standardizing the length of the data before entering into machine learning. One method used is TF-IDF. Term frequency - inverse document frequency (TF-IDF) is used to extract features from a document. With TF-IDF can convert text data into a matrix of numbers. The way TF-IDF works is to calculate how often or the frequency of occurrence of a particular word in the document [13]. Inverse document frequency measures the occurrence of any word in all documents. Then give the weight to each word. With TF-IDF it can also be used to normalize feature vectors.

TF-IDF is widely used for sentiment analysis, NLP and also text classification. Because the output of k-Mer is a substring and is included in the word or text. Then TF-IDF can be implemented in this DNA classification case.

2.3 Linear Discriminant Analysis (LDA)

LDA is the method used to reduce data dimensions. LDA is based on supervised learning which means it requires knowledge in it [9]. LDA work by calculating the linear discriminant between different classes and maximize the separation. The technique is compute inter and intra class distances [14]. Then homogeneous data will be close together, and heterogeneous data will be separated as far as possible. So, the data with its own class will gather together.

2.4 Support Vector Machine (SVM)

SVM is a machine learning method that can be used for classification and regression problems. In learning, SVM creates a hyperplane by maximizing data boundary margins between classes. Support vector is a predictor value that is closest to the border that separates the class. This support vector is used in calculating margin creation from SVM [9]. With this, SVM is able to generalize data to data that will be data. SVM can be applied to linear and non-linear problems using the kernel function. There are 4 kernels commonly used in SVM, namely linear, polynomial, sigmoid, and radial basis function (RBF). Linear kernels are used for data with a class that is liner separated. With this kernel it can run faster than other kernels, but when confronted with separate data that is not linear it will result in poor evaluation performance. The polynomial, sigmoid, and RBF kernels can be used for data with classes that are not linearly separated. For detailed about SVM is described in [15]. In this study the kernel used is RBF because it tends to run faster than the other kernels and is sometimes better at providing evaluation performance than SVM. In addition, cross validation will also be used with a fold = 10. Use cross validation to maximize classification performance.

2.5 Performance Measure

The performance of the machine learning classifier can be seen by its evaluation, such as accuracy, precision, recall, F score and Matthews Correlation Coefficient as outline below:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (4)$$

$$MCC = \frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

Where:

TP is True Positive, TN is True Negative
FP is False Positive, FN is False Negative

3 Result and Analysis

3.1 Feature Extraction

The used data is a complete genome of mycobacterium tuberculosis with 6 classes and a total of 233 data. The data class obtained from the lineage of the organism.

In this section, feature extraction is done using k-Mer then TF-IDF and the result shown as Table 1 below. The table shows data dimension from extraction feature process. The *k* column is the k-Mer parameter which shows the value of how many substrings DNA sequence to take. The result is the higher value of *k*, the higher dimension of data will be produced. It means the data need to be reduced dimension to lower dimension. It hopes to reduce computational time because with high dimension, it consumes high computational time

Table 1. Data dimension result using TF-IDF

k	Dimension
2	233, 198
3	233, 1533
4	233, 7576
5	233, 25602
6	233, 66321
7	233, 141495
8	233, 263926
9	233, 525081

After extracting the feature, next step is classification to obtain the evaluation by k-Mer and TF-IDF result shown at the Table 2 below.

Table 2. Evaluation classification using TF-IDF

k	2	3	4	5	6	7	8	9
Accuracy	0.661	0.661	0.661	0.661	0.661	0.661	0.661	0.661
Precision	0.627	0.627	0.627	0.627	0.627	0.627	0.627	0.627
Recall	0.661	0.661	0.661	0.661	0.661	0.661	0.661	0.661
F score	0.585	0.585	0.585	0.585	0.585	0.585	0.585	0.585
MCC	0.373	0.373	0.373	0.373	0.373	0.373	0.373	0.373
Time	0.217	0.893	2.44	3.68	9.95	30.8	109	385

From the table can be seen that the results of the classification give the same results for each *k* from the k-Mer data. With an average value of accuracy = 0.661, precision = 0.627, recall = 0.661, F score = 0.585 and MCC = 0.373.

At this stage, it is still not able to determine the best *k* value because the evaluation results of each *k* value there are no differences. Cause of that, by implementing dimension reduction it is expected to increase accuracy because based on previous research using dimensional reduction can improve the accuracy of its classification.

3.2 Dimension Reduction

After feature extraction, dimension reduction is necessary because the previous data has a large dimension enough for the higher *k* values. Also, the evaluation of classification before had no differences value. This dimension reduction aims to prove that dimensional reduction can change to lower dimension, reduce computational time and improve classification evaluation performance.

Table 3. Evaluation classification TF-IDF using LDA

k	2	3	4	5	6	7	8	9
Dimension	233, 5	233, 5	**233, 5**	233, 5	233, 5	233, 5	233, 5	233, 5
Accuracy	0.768	0.828	**0.927**	0.914	0.897	0.910	0.906	0.923
Precision	0.745	0.848	**0.930**	0.915	0.902	0.908	0.891	0.914
Recall	0.768	0.828	**0.927**	0.914	0.897	0.910	0.906	0.923
F score	0.740	0.824	**0.924**	0.909	0.891	0.905	0.895	0.913
MCC	0.600	0.718	**0.875**	0.852	0.823	0.842	0.834	0.866

Time	0.047	0.031	**0.047**	0.047	0.031	0.018	0.031	0.031

For the results of the experiment using LDA from the TF-IDF data it produces various evaluation values for each data with different *k*. Accuracy values tend to increase and decrease, and the most optimal is at k = 4 with an accuracy = 0.927, precision = 0.930, recall = 0.927, F score = 0.924, and MCC = 0.875 as shown in Table 3 above. Comparison accuracy before and after LDA applied can be seen in the Fig. 2 below.

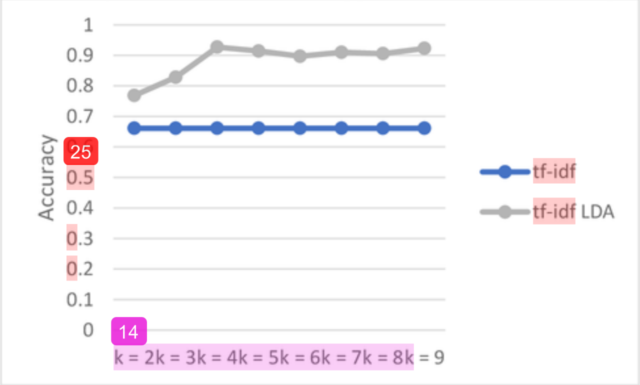

Figure 2. Accuracy comparison

Based on Fig. 2 above, the difference of accuracy is visible and accuracy is being better after the LDA is applied. We can see if the best *k* is *k* = 4 which give the best performance.

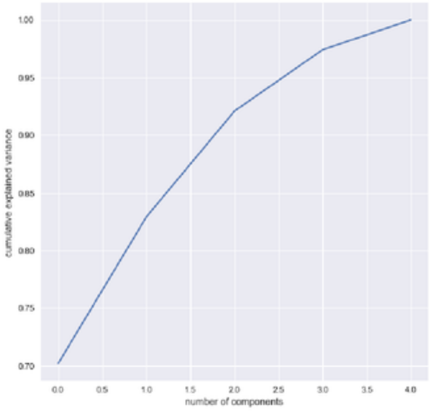

Figure 3. Cumulative explained variance from *k* = 4

Judging from the data dimensions generated using LDA, the data dimensions tend to be the same as 233, 5. In Fig. 3, an example of an explanation of the variant of LDA with k = 4. It can be seen with LDA producing dimensions 233, 5 because there are only 5 features that have the most influence from the data. We still can choose some features to determine what percentage of data after reduction will be used, but the reduced data has very small dimensions. So, there is no need for feature selection. Cumulative explain variance is the sum result of the explained variance ratio with the following values [0.70215354, 0.12704264, 0.09180755, 0.0531784, 0.025171788, 0.02581788].

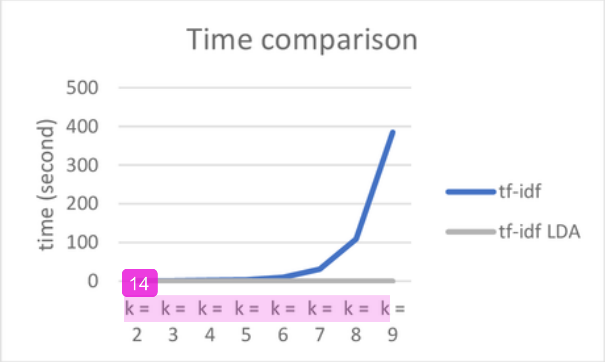

Figure 3. Time comparison with and without LDA

This dimension is smaller than before LDA applied and automatically the computing time needed is much faster. Whereas if observed from the evaluation side, the LDA provides better and more varied evaluations from the classifier for each different *k* value. This is because LD A is based on supervised learning which makes data tend to get closer to the same class and maximize boundaries between different classes. The average computing time from $k \geq 2 \leq 9$ is 0.0354 seconds. Taken averages because each *k* has the same dimension of data. From the average value, it can be seen that computing time is less than 1 second and is calculated to be much faster than without LDA.

It is proven that LDA can be used to reduce data dimensions and computation time, but can also improve the performance of the SVM classifier used in this study.

4 Conclusion

Based on the results of the study, it can be concluded that k-Mer has an influence on the extraction of Tuberculosis DNA data. The extraction of features used can determine the evaluation of classification. To extract data from DNA substring, we used TF-IDF and successfully applied. Likewise, with the selection of methods to reduce data dimensions. In this case, LDA is the best because the evaluation results are very good because the LDA is in the dimension reduction process, there is a learning process from the class that has been determined. Because of a feature reduction, computing time will automatically run faster than without the dimension reduction. The final result of this study is that the best *k* value is k = 4 based on the experiment using TF-IDF and LDA. With performance evaluation accuracy = 0.927, precision = 0.930, recall = 0.927, F score = 0.924, and MCC = 0.875.

References

[1] S. Asia, W. Paci, I. Congress, T. Evolution, and T. B. E. Meeting, "Tuberculosis in evolution," no. April, pp. 3–5, 2015.

[2] S. A. Yimer, G. Norheim, A. Namouchi, E. D. Zegeye, W. Kinander, and T. Tønjum, "Mycobacterium tuberculosis Lineage 7 Strains Are Associated with

Prolonged Patient Delay in Seeking Treatment for Pulmonary Tuberculosis in Amhara Region , Ethiopia," *J. Clin. Microbiol.*, vol. 53, no. 4, pp. 1301–1309, 2015.

[3] R. De Janeiro, "Artificial Neural Network Models for Diagnosis Support of Drug and Multidrug Resistant Tuberculosis," *Lat. Am. Congr. Comput. Intell.*, pp. 1–5, 2015.

[4] Y. Zhan, B. Li, Y. Huo, A. Lin, and H. Wu, "A case of multiple organ tuberculosis," *Radiol. Infect. Dis.*, pp. 0–4, 2018.

[5] J. T. Wassan, H. Wang, and H. Zheng, "Machine Learning in Bioinformatics," *Encycl. Bioinforma. Comput. Biol.*, pp. 300–308, 2019.

[6] W. Ashlock and S. Datta, "Evolved features for DNA sequence classification and their fitness landscapes," *IEEE Trans. Evol. Comput.*, vol. 17, no. 2, pp. 185–197, 2013.

[7] M. Martínez-porchas and F. Vargas-albores, "An efficient strategy using k-mers to analyse 16S rRNA sequences," *Heliyon*, no. May, p. e00370, 2017.

[8] G. Han and D. Cho, "Genomics Genome classification improvements based on k-mer intervals in sequences," *Genomics*, no. October, pp. 0–1, 2018.

[9] S. Ilias, N. Tahir, R. Jailani, and S. Alam, "Feature Extraction of Autism Gait Data Using Principal Component Analysis and Linear Discriminant Analysis," *2016 IEEE Ind. Electron. Appl. Conf.*, pp. 275–279, 2016.

[10] D. Novitasari, I. Cholissodin, and W. F. Mahmudy, "Optimizing SVR using Local Best PSO for Software Effort Estimation," *J. Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 28–37, 2016.

[11] D. Novitasari, I. Cholissodin, and W. F. Mahmudy, "Hybridizing PSO with SA for Optimizing SVR Applied to Software Effort Estimation," *TELKOMNIKA*, vol. 14, no. 1, pp. 245–253, 2016.

[12] D. Phan, N. G. Nguyen, F. R. Lumbanraja, and M. R. Faisal, "Combined Use of k-Mer Numerical Features and Position-Specific Categorical Features in Fixed-Length DNA Sequence Classification," *J. Biomed. Sci. Eng.*, vol. 10, no. 8, pp. 390–401, 2017.

[13] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, 2016.

[14] Y. Wang and Y. Chen, "A New Feature Extraction Algorithm Based on Fisher Linear Discriminant Analysis," *2017 3rd Int. Conf. Control. Autom. Robot.*, no. 1, pp. 414–417.

[15] V. N. Boser, Bernhard E. and Guyon, Isabelle M. and Vapnik, "Training Algorithm Margin for Optimal Classifiers," *COLT '92 Proc. fifth Annu. Work. Comput. Learn. theory*, pp. 144–152, 1992.

JITECS 113

ORIGINALITY REPORT

16%	14%	10%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jitecs.ub.ac.id Internet Source	2%
2	Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath. "Classification of sentiment reviews using n-gram machine learning approach", Expert Systems with Applications, 2016 Publication	1%
3	kanazawa-u.repo.nii.ac.jp Internet Source	1%
4	iptek.its.ac.id Internet Source	1%
5	www.fujipress.jp Internet Source	1%
6	www.irjet.net Internet Source	1%
7	www.cil.ntu.edu.sg Internet Source	1%
	journals.plos.org	

8	Internet Source	1%
9	Mulualem Tadesse, Gemeda Abebe, Alemayehu Bekele, Mesele Bezabih, Pim de Rijk, Conor J. Meehan, Bouke C. de Jong, Leen Rigouts. "The predominance of Ethiopian specific Mycobacterium tuberculosis families and minimal contribution of Mycobacterium bovis in tuberculous lymphadenitis patients in Southwest Ethiopia", Infection, Genetics and Evolution, 2017 Publication	1%
10	www.scielo.br Internet Source	1%
11	onlinelibrary.wiley.com Internet Source	1%
12	www.amalthea-reu.org Internet Source	1%
13	juti.if.its.ac.id Internet Source	<1%
14	www.numericana.com Internet Source	<1%
15	uitm.pure.elsevier.com Internet Source	<1%
16	www.sciencedaily.com	

	Internet Source	<1%
17	la-cci.org Internet Source	<1%
18	www.liebertpub.com Internet Source	<1%
19	Huiru Zheng, Jyotsna Talreja Wassan, Mihnea Alexandru Moisescu, Lacramioara Stoicu-Tivadar et al. "Multiscale Computing in Systems Medicine: a Brief Reflection", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018 Publication	<1%
20	export.arxiv.org Internet Source	<1%
21	L. H. R. A. Evora, J. M. Seixas, A. L. Kritski. "Artificial neural network models for diagnosis support of drug and multidrug resistant tuberculosis", 2015 Latin America Congress on Computational Intelligence (LA-CCI), 2015 Publication	<1%
22	www.kmice.cms.net.my Internet Source	<1%
23	www.un.org Internet Source	<1%

24	www.iosrjournals.org Internet Source	<1%
25	Liang Zheng, Shengjin Wang, Ziqiong Liu, Qi Tian. "Lp-Norm IDF for Large Scale Image Search", 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013 Publication	<1%
26	link.springer.com Internet Source	<1%
27	www.science.gov Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		