Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine

A Alamsyah* and T Fadila
Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang, Semarang, Indonesia

*alamsyah@mail.unnes.ac.id

Abstract. Data mining has been widely used to diagnose diseases from medical data. Classification is a data mining technique that can be used to predict disease. In previous studies, a support vector machine was widely used to obtain high accuracy in predicting hepatitis. In this study, the principal component analysis was applied to the support vector machine. A principal component analysis is used to extract features and reduce the number of features or attributes. Principal component analysis can reduce data dimensions without removing important information from the dataset. The extracted and reduced data are then used to classify the support vector machine. Classification performance measurement is done by using a confusion matrix. Hepatitis prediction accuracy achieved was 93.55%. This result is better than the support vector machine classification results without the application of principal component analysis.

1. Introduction
The development of artificial intelligence is very fast, including recurrent neural network used for forecasting [1], artificial intelligence neural networks are used to predict a disease [2,3] and support vector machines are used to improve the accuracy of predictions [4-6]. Nowadays, the healthcare industry has been generated large amounts of data and would be increased over time. This data volume increment requires an automatic way for these data to be extracted when needed [7]. The number of medical record data can be processed by utilizing data mining techniques. Data mining has been used widely in the field of bioinformatics to analyze biomedical data [8]. These medical data can be used by researchers to extract useful information by using data mining techniques such as classification techniques.

Classification is a technique that can be used to predict data or describe data classes. The classification algorithm can be used to help medical experts to diagnose a disease. In this study, the classification will be applied to predict hepatitis patient data. Determining the diagnosis of hepatitis is a difficult task for physicians because it has so many factors that should be considered and analyzed [9,10].

One of the popular classification algorithms used to make predictions is the support vector machine (SVM). the SVM is known to work well on high dimensional data. Abdullah et al., [11] applied Support Vector Machine (SVM), multilayer perceptron (MLP) dan K-nearest neighbor (KNN) to improve hepatitis classification. They analyzed the hepatitis patients from the UC Irvine machine learning repository. They concluded that KNN the best classifier to diagnose hepatitis disease is KNN, which
gives a higher degree of accuracy up to 100%. Both MLP and SVM showed their efficiency compared to previous work with the obtained accuracy is 97.87%.

Anto and Candramathi [12] proposes a medical decision support system that uses SVM for classification and hybrid Genetic Algorithm - Simulated Annealing for optimization. The performance of the expert system is analyzed using various parameters like classification accuracy, sensitivity, and specificity. Using the 80-20 train-test combination, the proposed methods are constructed and tested. The obtained results were 87% of accuracy, 82% of specificity, and 93% for sensitivity.

However, the SVM requires large memory if the processed data has high dimensions or has many attributes [13]. Therefore, the principal components analysis (PCA) method is applied to reduce the dimensions of the dataset. Many data mining algorithms work better if the dimensions of the dataset are lower, this is because dimensional reduction can remove irrelevant features, reduce noise. The PCA can reduce the dimensions of data by combining the essence of attributes by forming alternatives of smaller feature subset.

This study has purposed a combination of support vector machine (SVM) and principal components analysis (PCA) method. The SVM was chosen because it can produce high accuracy, and the PCA is used to extract the best features of the dataset and used to reduce the number of features to obtain better results.

2. Methods
In this study, a combination of principal components analysis and support vector machine is applied to predict the hepatitis dataset. The PCA is used to extract the hepatitis dataset and reduce the number of dataset features. The SVM algorithm is used to classify the reduced data. The classification results are evaluated based on the confusion matrix by calculating the accuracy of the model. The flowchart of the method used in this study is shown in Figure 1.

![Flowchart of the PCA-SVM method](image)

**Figure 1.** Flowchart of the PCA-SVM method.
2.1. Data preprocessing

Data preprocessing is an important step in the data mining process because the data collected is usually incomplete, redundant, noisy, and inconsistent. It is important to prepare appropriate data for the data mining process to obtain optimal results. In this study, data preprocessing was carried out in several steps, including cleaning data by handling missing values, data transformation by performing standardization or normalization, and data reduction using the PCA method. The used data is obtained from the UCI machine learning repository. The hepatitis dataset has 19 attributes and 155 samples.

The hepatitis dataset contains 167 missing values or 5.7% of the entire dataset. Missing value can reduce the level of accuracy of the data and the quality of the data will decrease in further data processing, such as the classification process. These missing values need to be handled to obtain a better classification result. Handling missing values was done by replacing the missing values using the most frequent value (mode).

2.2. Principal component analysis

Principal Component Analysis (PCA) is a technique to construct new variables from linear combinations of the original variables. This new variable is also known as the principal component (PC). The maximum number of new variables (PC) is equal to the number of original variables.

The stages of the Principal Component Analysis method are as follows:

2.2.1. Standardize the data by calculating the Z-score. Data standardization or normalization is done by calculating the Z-score based on the mean and standard deviation of each attribute. The Z-score value can be calculated by Equation 1.

\[ Z = \frac{x_j - \bar{x}_j}{\sigma} \]  

\( Z \) = Z-score value  
\( x_i \) = Data sample  
\( \bar{x} \) = Mean of attribute  
\( \sigma \) = Standard Deviation

Standard Deviation can be calculated by Equation 2.

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \]  

2.2.2. Calculate the covariance matrix by Equation 3.

\[ Cov = \frac{(x_i - \bar{x})^T(x_i - \bar{x})}{n-1} \]  

2.2.3. Calculate the eigenvalue and eigenvector of the covariance matrix based on Equation 4.

\( \det[A - \lambda I] = 0 \)  
\( A = \) Covariance matrix  
\( \lambda = \) Eigenvalue  
\( I = \) Identity matrix

Eigenvector can be calculated using equation 5.

\[ [A - \lambda I]X = 0 \]  

\( X \) is an eigenvector of each eigenvalue.

2.2.4. Choosing components and forming a feature vector. After the eigenvectors are formed from the covariance matrix, the next step is to order them by eigenvalue from highest to lowest. The feature vector
is constructed by taking the eigenvectors that you want to keep from the list of eigenvectors in order of significance.

2.2.5. **Deriving the new data set by multiplying the transpose of the feature vector with the transpose of the standardized dataset.**

2.3. **Support vector machine**

The stages of the Support Vector Machine algorithm in data classification [14-15] are as follows:

2.3.1. **Prepare training data.** In this study, the data is split into two, 80% for training data and 20% for testing data.

2.3.2. **Finding boundaries between classes.** When each point in a class is connected to another point, a line that separates between the classes will appear. This limit is known as the convex hull. Each class has its convex hull and because the class (assumed) is linearly separated, this hull does not intersect.

2.3.3. **Determine a hyperplane that maximizes the margin between classes.** Can be done in the following ways:

   a. First, any hyperplane stated in two attributes, \(x_1\) and \(x_2\), can be written in Equation 6.

      \[
      w \cdot x + b = 0
      \]  

   \[w = \text{weight} (w_1, w_2, \ldots, w_n)\]  

   \[x = \text{number of attribute} (x_1, x_2, \ldots, x_n)\]  

   \[b = \text{bias}\]  

   After defining a hyperplane in this mode, margin can be written in equation 8.

   \[\text{margin} = \frac{2}{\sqrt{w_0}} \cdot w_0\]  

   b. Maximizing this quantity requires quadratic programming, which is a process that has a strong position in the theory of mathematical optimization. Furthermore, we can be easily stated in some examples of training data, known as support vectors, which can be written in equation 9.

   \[|w_0| = \sum y_i x_i\]  

   \[y_i = \text{class label (+1 and -1 for biner classification)}\]  

   \[x_i = \text{support vector}\]  

   \[i = \text{a zero coefficient only for the support vector}\]  

2.3.4. **After setting boundaries and hyperplane, each new test can be classified by calculating on which side of the data results in the hyperplane.** This can be found by replacing the test \(x\) example into the hyperplane equation. If you count +1, then it includes a positive class and if it is calculated as -1, then it belongs to the negative class.

2.4. **Evaluation**

The evaluation of the classification model is based on testing to predict true and false objects, the measurement of classification performance is done with the confusion matrix.

The confusion matrix method is used to calculate the accuracy obtained in the classification process. Accuracy is the percentage of the total data classified correctly. Measuring accuracy can be written in Equation 10.

\[
\text{Accuracy} = \frac{TP+TN}{P+N} \times 100\%
\]  

The error rate is also calculated in the classification. Measuring the level of error can be done by Equation 11.

\[
\text{Error Rate} = \frac{FP+FN}{P+N} \times 100\%
\]
3. Results and discussion
In this study, the hepatitis dataset is divided into two, training data contains 124 samples (80%) and testing data contains 31 samples (20%). In outline, the research consists of two stages, i.e. the extraction and feature reduction stages using Principal Component Analysis and the classification stage using the Support Vector Machine algorithm.

3.1. Feature extraction and reduction
Principal Component Analysis (PCA) method is applied to extract and reduce features of the hepatitis dataset by forming a new dataset from a linear combination of the original features. From 19 attributes of hepatitis formed 8 new features called the principal component (PC). The PC contains an explained variance of approximately 70% or it can be said to represent 70% of the information from the entire dataset. The formed attributes were then called as PC1, PC2, PC3, PC4, PC5, PC6, PC7, and PC8. From each PC1 up to PC8 had a variance of 22.07%, 10.26%, 8.29%, 7.60%, 6.40%, 5.42%, 5.01%, and 4.98%. Hence the total explained variance is 70.03%.

3.2. Classification
Data classification was carried out by 2 scenarios. The first scenario is to classify the hepatitis dataset using only the SVM algorithm and the second scenario performs classification using a combination of the SVM and the PCA algorithms as a feature reduction on the standardized hepatitis dataset. From the first scenario, the implementation of the SVM algorithm results in 87.1% accuracy based on the calculation of the confusion matrix. From the results of the SVM-PCA model implementation, the accuracy level obtained was 93.55%. The comparison of the level of accuracy obtained by the SVM algorithm and the combination of the SVM-the PCA is shown in Figure 2.

![Figure 2. The comparison of the level of accuracy.](image)

According to Figure 2, it can be seen that the accuracy of the SVM classifier has increased by 6.45% after the PCA is applied to the dataset. The increase in accuracy results might be caused by centralized data and then the data is scaled into a smaller space. On the other side, the PCA can prioritize features based on the sign so that features with low variance or irrelevant can be eliminated, hence the SVM algorithm has better performance.

4. Conclusion
Based on the results and discussion, it can be concluded that the implementation of the PCA algorithm in the hepatitis dataset classification can reduce the dimensions or features of the dataset by forming a new dataset. Initially, 19 attributes can be reduced to 8 attributes with a fixed number of instances, i.e.
The classification result by applying the SVM algorithm without preprocessing is accuracy of 87.01%. While the classification results by applying the SVM algorithm using the PCA dataset is 93.54% of accuracy. Thus, it can be said that the application of the PCA in the SVM classification algorithm can predict hepatitis with good accuracy of 93.54%.

References

[1] Walid and Alamsyah 2017 Recurrent Neural Network For Forecasting Time Series With Long Memory Pattern J. Phys.: Conf. Ser. 824 012038
[2] Vijayashree J and Sultana H P 2018 A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier Programming and Computer Software 44(6) 388–397
[3] Sonobe T, Tabuchi H, Ohsugi H, Masumoto H, Ishitobi N, Morita S, Enno H, and Nagasato D 2019 Comparison between support vector machine and deep learning, machine-learning technologies for detecting epiretinal membrane using 3D-OCT Int Ophthalmol 39 1871–1877
[4] Lu Y and Roychowdhury V 2008 Parallel randomized sampling for support vector machine (SVM) and support vector regression (SVR) Knowl Inf Syst 14 233–247
[5] Wenqing N I U, Yinaer H A and Nan C H 2020 Support vector machine based machine learning method for GS 8QAM constellation classification seamless integrated fiber and visible light communication system SCIENCE CHINA Information Sciences 63(10) 1-12
[6] Abdolrazzaghi M, Hashemy S, and Abdolali A 2018 Fast-forward solver for inhomogeneous media using machine learning methods: artificial neural network, support vector machine and fuzzy logic Neural Comput & Applic 29 1583–1591
[7] Milovic B, and Milovic M 2012 Prediction and decision making in Health Care using Data Mining. International Journal of Public Health Science (IJPHS) 1 69–76
[8] Hussien S O, Elkhatem S S, Osman N and Ibrahim A O 2018 Proceedings of 2017 Sudan Conference on Computer Science and Information Technology 1–6
[9] Polat K and Güneş S 2006 Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. Digital Signal Processing: A Review Journal 16 889–901
[10] Nilashi M, Ahmadi H, Shahmoradi L, Ibrahim O, and Akbari E 2019 A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique Journal of Infection and Public Health 12 13–20
[11] Abdullah H N, Abd B H and Muhi S H 2018 High-Resolution Systems for Automated Diagnosis of Hepatitis 2018 Third Scientific Conference of Electrical Engineering SCEE 39–44
[12] S Anto and S Chandramathi 2015 An Expert System based on SVM and Hybrid GA-SA Optimization for Hepatitis Diagnosis International Journal Of Computer Engineering In Research Trends 2 437-43
[13] Kotu V and Deshpande B 2015. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner (Waltham: Elsevier)
[14] Han J, Kamber M and Pei J 2011 Data Mining: Concepts and Techniques 3rd ed (Waltham: Elsevier)
[15] Sutarti S, Putra A T and Sugiharti E 2019 Comparison of PCA and 2DPCA Accuracy with K-Nearest Neighbor Classification in Face Image Recognition Scientific Journal of Informatics 6 64–72