Support vector machine and principal component analysis for microarray data classification

Widi Astuti\textsuperscript{1}, Adiwijaya\textsuperscript{2}

School of Computing, Telkom University, Bandung

\textsuperscript{1}astutiwidi@telkomuniversity.ac.id; \textsuperscript{2}adiwijaya@telkomuniversity.ac.id;

Abstract. Cancer is a leading cause of death worldwide although a significant proportion of it can be cured if it is detected early. In recent decades, technology called microarray takes an important role in the diagnosis of cancer. By using data mining technique, microarray data classification can be performed to improve the accuracy of cancer diagnosis compared to traditional techniques. The characteristic of microarray data is small sample but it has huge dimension. Since that, there is a challenge for researcher to provide solutions for microarray data classification with high performance in both accuracy and running time. This research proposed the usage of Principal Component Analysis (PCA) as a dimension reduction method along with Support Vector Method (SVM) optimized by kernel functions as a classifier for microarray data classification. The proposed scheme was applied on seven data sets using 5-fold cross validation and then evaluation and analysis conducted on term of both accuracy and running time. The result showed that the scheme can obtained 100% accuracy for Ovarian and Lung Cancer data when Linear and Cubic kernel functions are used. In term of running time, PCA greatly reduced the running time for every data sets.

1. Introduction
Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Based on WHO cancer fact sheets [1], cancer is a leading cause of death worldwide, accounting for 8.2 million deaths in 2012. Conventional methods for monitoring and diagnosing cancers there are now very dependent on human observation to detect certain features. A cancer diagnosis is usually performed using imaging systems and analysis of morphological and clinical data. In recent decades Microarray take an important role in the diagnosis of cancer and improve the accuracy of cancer diagnosis compared to traditional techniques. With microarrays, it is possible to examine a gene expression within a single sample or to compare gene expressions within two tissue samples, such as in tumor and non-tumor tissues [2].

The characteristic of microarray data is small sample but huge dimension. Since that, there was a challenge for researcher to provide solutions for microarray data classification with high performance in both accuracy and running time. Some researches on cancer detection using microarray data are listed as follow. In 2007, Gavin and Talbot [3] proposed a modification for feature selection using sparse logistic regression [4] with Bayesian regularization which considered fast compared to previous method but the error rate is 17%. In 2012, Sahu et al [5] proposed a new approach for microarray data classification using feature selection which utilized Particle Swarm Optimization (PSO). The research showed that the usage of PSO on feature selection process could increase the accuracy. In 2014, Tran et all [6] did a review on PSO-based feature selection for data classification. Tran remarked that
feature selection is a NP hard problem which is very challenging especially for high dimensional data. The use of PSO or another Evolutionary Computing method requires precision in represent the data into the searching space. Furthermore, Tran stated that on feature selection there is trade off conflict between increasing accuracy and reducing computational cost.

In 2015, Vanitha [7] proposed SVM along with mutual information gain for its feature selection. Vanitha used four different SVM namely linear SVM, SVM with RBF kernel, SVM with quadratic kernel, and SVM with polynomial kernel. Furthermore, Vanitha compared the proposed scheme result with KNN and ANN. Based on the result, the usage of MI-SVM gave better result than KNN and ANN and even hit 100% accuracy for some cases. The performance is low for colon cancer data set with 74.19% as its highest accuracy which is achieved from linear SVM and 38.70% accuracy achieved from SVM with quadratic kernel. Nurfalah [8] used PCA as feature selection for ANN based microarray data classification. Mollae [8] combined three feature selection methods and a feature extraction method called PSO-dICA before entering the SVM process. Mollae suggested that in future works, kernel functions are used. This research attempted to improve Vanitha [7] result and adopted [9] suggestion for better performance.

2. Microarray Data Classification
In our work, there are generally three steps which is illustrated on figure 1. First step is feature extraction using PCA. Next is classification using SVM with different kernels. Last step is performance evaluation using accuracy measurement.

2.1. Feature Extraction
As the dimension (number of features) increase, it will be harder to do the data mining task such as classification. To solve that problem, Feature Extraction namely PCA which extract original features into a new features using mapping function is used. PCA processes include: data centering, calculate covariance matrix, calculate Eigenvector and Eigenvalue, select top Eigenvector, and transformation data.

2.2. Classification
Support Vector Machine (SVM) is well known as a classifier which can model complex data, has good accuracy and less prone to overfitting. SVM works by searching the linear optimal separating hyperplane (decision boundary). The rationale is that decision boundary with large margin is better when handling unseen data compared to decision boundary with small margin. When the data are not linearly separable, SVM transform original data into a higher dimension using a nonlinear mapping to obtain the separating hyperplane.
2.3. Performance Evaluation
In this study, there are two type of experiments conducted. First experiment type is conducted to see the effect of the number of component chosen in PCA. The first experiment will be conducted on colon cancer data set. Colon cancer data set is chosen due to its position as performance comparison with previous research. Second experiment is conducted to see the proposed method’s accuracies and running time. The experiment is conducted on seven data set, namely: Breast Cancer, Nervous System Cancer, Colon Cancer, Lung Cancer, Leukemia, Ovarian Cancer, and Prostate Cancer data. For each data set, 12 scenarios of experiment which explained on Table 1 will be performed.

| Scenario # | Reduction Dimension | Kernel in SVM     |
|------------|---------------------|-------------------|
| 1          | None                | Linear            |
| 2          | None                | Quadratic         |
| 3          | None                | Cubic             |
| 4          | None                | Fine Gaussian     |
| 5          | None                | Medium Gaussian   |
| 6          | None                | Coarse Gaussian   |
| 7          | PCA                 | Linear            |
| 8          | PCA                 | Quadratic         |
| 9          | PCA                 | Cubic             |
| 10         | PCA                 | Fine Gaussian     |
| 11         | PCA                 | Medium Gaussian   |
| 12         | PCA                 | Coarse Gaussian   |
Microarray data that have very small sample can complicate the evaluation to generated model so that evaluation results can over fit to only one subset of data. We apply the k-fold cross validation technique to resolve this problem, k-fold cross validation technique ensure that evaluation of the resulting model is not over fit to only one data so that model selection is more valid. By using k-fold cross validation, model selected and evaluated through k times iteration. The diagram of experiment process in this study which using k-fold cross validation can be seen in Figure 2. The number of k which are used in this study is 5.
3. Testing and Analysis
As explained on 2.3, an experiment on Colon Cancer Data is conducted to see the effect of the number of component chosen in PCA. Colon Cancer Data is chosen due to its position as performance comparison with previous research. Colon Cancer Data contains 62 samples with 2,000 features. After the explained variances are discovered, the colon cancer data is transformed into new data set which has lower dimension than original data. Nine different experiment scenarios are conducted and the accuracies are measured for six different SVM. The nine scenarios are chosen using explained variance threshold. The nine thresholds are 25%, 45%, 65%, 80%, 90%, 95%, 97.5%, 98.75% and 100%. It means first scenario would pick a minimum number of component which satisfies variance kept ≥ 25%, and so on. After the PCA conducted, the results are put into six different SVMs. The accuracies are measured for each classification results. The accuracies are presented on Table 2.

From Table 2, it can be seen that for Linear SVM, the best accuracy obtained when the number of component chosen is 8 or 16. From Quadratic and Cubic SVM, best accuracy is obtained when the number of component is 25 which keep 95% variance during PCA process. Best accuracy for Coarse Gaussian is obtained when the number of component after PCA is 43. Furthermore, best accuracy for Fine Gaussian and Medium Gaussian is 67.7% which is obtained when the number of component kept is 4 and 43 respectively.

From accuracy average for each different scenarios, it can be seen that the higher explained variance, the accuracy also increases. This trend stopped on 95% threshold and decreases after that. Best result of all scenario of choosing number of component and 6 different SVM obtained by Quadratic SVM with 95% variance kept after PCA.

The experiment result for second experiment is presented on Table 3 and Table 4. Table 3 shows the accuracies of each scenarios for each data set while Table 4 shows the running time of each scenarios for each data set.

| Number of Components after PCA | Variance Kept after PCA (%) | Accuracy (%) |
|-------------------------------|-----------------------------|--------------|
| 1                             | 36.1 48.44 66 80.14 90.34 95.08 97.52 98.85 100 |
| 2                             | Linear 64.5 66.1 75.8 83.9 83.9 79.0 72.6 69.4 66.1 |
|                               | Quadratic 62.9 62.9 74.2 79.0 82.3 85.5 80.6 79.0 75.8 |
|                               | Cubic 45.2 72.6 79.0 72.6 79.0 82.3 79.0 80.6 75.8 |
|                               | Fine Gaussian 59.7 67.7 67.7 64.5 64.5 62.9 64.5 64.5 64.5 |
|                               | Medium Gaussian 62.9 66.1 66.1 64.5 62.9 64.5 64.5 67.7 64.5 |
|                               | Coarse Gaussian 64.5 64.5 62.9 62.9 64.5 64.5 64.5 64.5 64.5 |
|                               | Average 59.95 66.65 70.95 71.23 72.85 73.12 70.95 70.95 68.53 |

Table 3. Accuracies for proposed method.

| Scenario # | Breast | Nervous | Colon | Lung | Ovarian | Leukemia | Prostate |
|------------|--------|---------|-------|------|---------|----------|----------|
| 1          | 56.41  | 61.70   | 83.90 | 100.00 | 100.00  | 87.70    | 93.10    |
| 2          | 65.41  | 61.70   | 85.50 | 100.00 | 99.60   | 87.70    | 92.20    |
For breast cancer data, it can be seen that accuracy from scenario without using PCA is better than scenario with PCA process, especially for quadratic and cubic kernel, but relatively similar for others kernel. For Nervous cancer, surprisingly accuracy from scenario which used PCA is better than scenario without PCA process in 3 kernels, and similar in other kernels. Accuracy for colon data from scenario without dimension reduction is better for SVM with Medium Gaussian Kernel, but relatively similar for other kernels. In lung cancer case, the accuracy of the scenario without dimensional reduction process is slightly larger for SVM with linear, quadratic, cubic, and Gaussian kernels. However, the accuracy of the scenario using PCA is better for SVM with the Gaussian and Coarse Gaussian kernels. Furthermore, the mean of scenario accuracy using PCA is greater than the mean of scenario accuracy without using PCA. The accuracy of the scenario without dimensional reduction process is better for SVM with the Gaussian Medium and Coarse Gaussian kernels, but relatively similar for other kernel types in ovarian cancer data. For leukemia data, in terms of accuracy, scenarios using dimensional reduction processes result in higher performance than scenarios without dimensional reduction for 5 SVM kernels other than the Medium Guassian kernel. Lastly, for prostate cancer data, accuracy of scenarios without dimensional reduction processes is generally higher than scenarios using PCA. In term of running time, scenarios with PCA constantly give better running time.

**Table 4.** Running Times for proposed method.

| Scenario # | Breast | Nervous | Colon | Lung | Ovarian | Leukemia | Prostate |
|------------|--------|---------|-------|------|---------|----------|----------|
| 1          | 520.32 | 50.40   | 8.99  | 103.07 | 157.41  | 102.43   | 101.45   |
| 2          | 493.48 | 52.13   | 5.64  | 97.74 | 157.03  | 98.98    | 99.73    |
| 3          | 493.62 | 46.80   | 5.55  | 96.75 | 156.61  | 95.65    | 99.34    |
| 4          | 357.08 | 48.37   | 5.63  | 96.83 | 159.91  | 95.96    | 96.84    |
| 5          | 357.08 | 48.73   | 5.40  | 96.48 | 157.52  | 99.06    | 101.52   |
| 6          | 356.44 | 53.20   | 5.423 | 97.66 | 158.71  | 122.96   | 100.02   |
| 7          | 164.10 | 29.89   | 8.10  | 54.36 | 80.327  | 62.32    | 58.10    |
| 8          | 162.80 | 27.88   | 7.23  | 49.24 | 81.794  | 60.32    | 51.41    |
| 9          | 162.39 | 28.15   | 7.04  | 48.82 | 162.39  | 51.63    | 69.85    |
| 10         | 147.13 | 28.00   | 7.20  | 49.98 | 147.13  | 58.63    | 50.58    |
| 11         | 155.82 | 27.85   | 7.17  | 48.87 | 155.82  | 51.52    | 57.68    |
| 12         | 146.54 | 27.68   | 7.09  | 49.04 | 146.54  | 51.14    | 50.69    |
4. Conclusion

From experiment result, it can be concluded that proposed scheme for microarray data classification which utilize Principal Component Analysis as dimension reduction method and SVM with kernel functions as classifier can surpass performance of previous research [7]. The usage of Principal Component Analysis proven to reduce the running time. Compared to [7] which also used colon cancer data set, the proposed scheme which utilizes PCA produces better accuracy by 4.81% for Linear SVM and 46.80% for Quadratic SVM. Various kernel function had been explored in this research along with suggestion from [9]. The result showed that from 6 from 7 highest accuracy for each data are obtained when kernel function is not used except for Breast Cancer Data when Quadratic Kernel Function increase the accuracy by 1.6%. We can concluded that the microarray data which had been used are linearly separable.

References

[1] World Health Organization, Cancer fact sheet, Retrieved from : http://www.who.int/mediacentre/factsheets/fs297/en/, Accessed : February 2016.

[2] Sarhan Ahmad M., Cancer Classification Based on Microarray Gene Expression Data Using DCT and ANN, Journal of Theoretical and Applied Information Technology, 2009.

[3] Cawley GC, Nicola L, Talbot C. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics 2006;22(19).

[4] Shevade S.K., Keerthi S.S.. A simple and efficient algorithm for gene selection using sparse logistic regression, Bioinformatics,2003, vol. 19.

[5] Sahu B., Mishra D., A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data, Procedia Engineering, 2012 (38).

[6] Tran B., Xue B., Zhang M. (2014) Overview of Particle Swarm Optimisation for Feature Selection in Classification. In: Dick G. et al. (eds) Simulated Evolution and Learning. SEAL 2014. Lecture Notes in Computer Science, vol 8886. Springer, Cham

[7] Vanitha, C.D.A., Devaraj, D. and Venkatesulu, M.. 2015. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science, 47, pp.13-21.

[8] Nurfalah A., Adiwijaya, Suryani A.S.. 2016. Far East Journal of Electronics and Communications. 16 (2). pp. 269-281.

[9] Mollaee, M., & Moattar, M. H. (2016). A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybernetics and Biomedical Engineering, 36(3), 521-529