Recursive Particle Swarm Optimization (RPSO) schemed Support Vector Machine (SVM) Implementation for Microarray Data Analysis on Chronic Kidney Disease (CKD)

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052077 doi:10.1088/1757-899X/546/5/052077

Recursive Particle Swarm Optimization (RPSO) schemed Support Vector Machine (SVM) Implementation for Microarray Data Analysis on Chronic Kidney Disease (CKD)

Zuherman Rustam[1*], Mas Andam Syarifah[1] and Titin Siswantining[1]

[1]Departemen of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA) Universitas Indonesia, Depok 16424, Indonesia

*Corresponding Author: zuherman@ui.ac.id

Abstract. Chronic Kidney Disease is the second chronical and catastrophic disease after heart disease in terms of treatment cost. This is because CKD symptoms occurs on final stages, that is fourth and fifth, in which it is too late for treatment. Therefore, final stage patient must receive continuous medication, such as haemodialysis. So early detection on a patient CKD is necessary to prevent patient to be chronic. Studies of gene genes are used to classify microarray data with global CKD decisions or not. So to get accurate results in this study using SVM-RFE with the addition of the Particle Swarm Optimization algorithm as a gene selector to be more optimal and it consideration of the fixed gene in its condition which is important information of the CKD gene itself. This research is then expected to be able to classify globally with CKD output or not CKD. As a result, for the CKD microarray data accuracy using RPSO schemed SVM highest than only using SVM-RFE.
Keyword: Chronic Kidney Disease, Classification, Feature Selection, SVM-RFE, RPSO, SVM

1. Introduction

Machine learning is now a trendy tool for researchers regarding Big data. Solving problems related to big data is very common now. Large data if done manually is not very efficient in terms of energy time and maybe even the accuracy or error that occurs will be higher [1] is the same as in classification cases involving a lot of data.

Cancer or chronic disease is the second largest cause of death in the world, because it is widely used as research material in the health sector continuously. CKD or Chronic Kidney Disease is one of the chronic diseases also classified as a catastrophic disease after heart disease in Indonesia. Even CKD is also a worldwide public health problem [3]. There are 5 stages of CKD, 1-3 is an early stage where hope of recovery is still high, while 4-5 is a final stage where most parts of the kidneys are not functioning, and patients must be treated with proper handling [2].

Because of late diagnosis, or improper treatment. need preventive measures that can diagnose right before getting worse, in accordance with the previous statement [2], it is expected that with early detection and appropriate treatment the patient will get faster recovery and higher patient life expectancy. There have been many studies regarding the classification of CKD data sets.

The current gene data microarray is a tool that can profiling gene expression and has been proven to be a value that determines the classification of complex diseases such as cancer and chronic diseases. Microarray information can also be used as the right treatment decision for patients [4]. However, the data generated by this microarray has hundreds, even tens of thousands of features. This feature is

explained about the genes themselves. This, of course, if processed directly, it will take time, energy and even a small amount of money. Accuracy in decisions can also be inaccurate because not all genes influence the classification decisions. Therefore, feature selection that can choose genes that are influential requires that the feature selection process becomes a fundamental stage to obtain high-dimensional data analysis [5].

Currently the research has been combined with the best optimum solution. In this study combining PSO with SVM-RFE to get the best feature subset by determining optimized parameters and then expected to optimize the classifier. The classification methods used are SVM, which is the oldest machine learning classifier, but very powerful because it can produce high accuracy compared to other classifiers. SVM development has now been combined with the kernel which can help resolve classifications in higher dimensions to complete non-linear classifications.

2. Methods

Many feature selection methods are currently being developed, because in processing big data using all features will consume time, memory capacities and costs. Therefore, many ways have been done to optimize good feature selection. One of them is by using an optimization method in determining the best parameters in the gen selection algorithm [7]. In this research we will use optimization in classification with SVM, namely SVM-RFE with the addition of PSO or Particle Swarm Optimization.

2.1. Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a fairly well-known optimization method in which this method is inspired by a group of animals that feed in groups by dividing and updating information for each individual groups to find the optimal solution. This makes the researcher get an idea where for each particle with the best value will continue to be updated according to the position of each particle [7]. PSO works to find the value of each particle flying experience (pBest) and companions experience (gBest). Each particle has a value that will be evaluated by the fitness function to be optimal, position and velocity that control the movement of each particle [9].

Table 1. Particle Swarm Optimization.

Particle Swarm Optimization	
Step 1	initialization of PSO parameters
Step 2	compute the cost of each particle using the fitness function
Step 3	seach the $pBest$ and $gBest$ values
Step 4	update the particle velocity v_i : $v_i\ (t) = wv_i(t-1) + c_1r_1\left(x_{pBest}(t) - x_i(t)\right) + c_2r_2\left(x_{pBest}(t) - x_i(t)\right) \dots (1)$
Step 5	update the particle position x_i : $x_i(t+1) = x_i(t) + v_i(t) \dots (2)$

2.2. SVM – RFE

Support Vector Machines – Recursive Feature Elimination as a feature selection that can eliminated feature repeatedly the work system selects features with the smallest values. By using the SVM algorithm and the concept of the RFE we can eliminate the features that have the smallest value in each iteration [7].

Table 2. Support Vector Machine – Recursive Feature Elimination.

Support Vector Machine – Recursive Feature Elimination	
Step 1	Initialization the original feature
Step 2	Loop the following procedure until $R = \emptyset$
Step 3	Obtain the training set with candidate feature set
Step 4	Train the SVM classifier to get *w*
Step 5	Calculate the ranking criteria score; $c_k = w_k^2, k = 1,2,\dots,\lvert S\rvert$ … (3)
Step 6	Find the smallest ranking criteria score features; $\arg\min_k c_k$ … (4)
Step 7	Update Feature set $R = P \cup R$ … (5)
Step 8	Remove this feature in S such thar $S = S/P$ (6)

2.3. Classification Methods

SVM that used in this research is SVM classification to differentiate the classifications [5]

$$\min_{w,b} \frac{\|w\|}{2}$$

$$s.t \qquad y^{(!)}\left(w^T x^{(i)} + b\right) \geq \in \{1, 2, \dots, N\} \dots (7)$$

SVM or Support Vector Machines as classifiers although they conventional, but their utilization and development are still being carried out because the idea of SVM and its accuracy results are still good in many classifications. Not only maximizing margins but also minimizing existing errors. SVM development used in this research is SVM with kernel trick using the kernel function, namely $K(x_i, x_j)$ $=\Phi(x_i)^T \Phi(x_j)$, which helps us to calculate higher dimensions without having to work directly on its dimensions. The higher dimension $\Phi(x_j) \in H$, which is called the feature space. The equation of the new classification function is obtained as follows:

$$f(x) = sgn(\sum_{i=1}^{N} y_i a_i \,.\, K\left(x_i, x_j\right) + b \dots (8)$$

The type of kernels that this research use is kernel polynomial and linear kernel to compare the accuracy both classification process with SVM. Polynomial kernel function and linear is defined as[5]:

1. $K(x_i, x_j) = (x_i, x_j + 1)^d$, where *d* represent the degree of Polynomial Kernel Function. … (9)
2. $y^{(!)}\left(w^T x^{(i)} + b\right) \geq \in \{1, 2, \dots, N\} \dots (10)$

2.4. Evaluation Method

Evaluation method in this classification for see the performance of the models in this research is used performance evaluation. Performance evaluation is step that can see significant of the model. The evaluation is performed accuracy, precision, and recall. The performed measuring by the value of True Positive (TP) which indicates the positive data entered into the system is detected correctly by the system, True Negative (TN) indicates negative data entered into the system is detected incorrectly by the system False Positive (FP) indicates the negative data entered into the system is detected correctly by the system, and False Negative (FN) indicates positive data entered into the system is detected incorrectly by the system.

They will be used to evaluate the performance of the proposed technique. The measures are computed using the following equation [10]:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{\text{CKD diagnosed properly}}{total\ samples} \quad \ldots (11)$$

$$Precision = \frac{TP}{TP+FP} \quad \ldots (12)$$

$$Recall = \frac{TP}{TP+FN} \quad \ldots (13)$$

3. Result and Discussion

3.1. Microarray dataset of gene expression CKD

This research use GSE97709 dataset microarray on NCBI. Plasma transcriptome profiling in patients with chronic kidney disease and the url address can be accessed for free on the following page *https://www.ncbi.nlm.nih.gov /geo/query/acc.cgi?acc=GSE97709*. Data set includes 159 samples, 48 training and testing data samples 22 control samples and 26 CKD samples. 111 sample as validation data, 13 controls and 98 CKD samples. Data set division can be seen in Table 1.

Table 3 Division of DNA microarray data on GEO NCBI.

	Health	CKD	Total
Training Set	5	30	35
Testing Set	8	5	13
Total	13	35	48

3.2. Experiment Result

The experimental results for this microarray data are obtained by representing data in the form of matrices with sizes m x n with rows representing samples and columns as features or in this research genes. Each element value in the matrix is expression level of gene, which will then be extracted features, feature selection and classification to it.

In this research, the first extraction feature is done to make the desired matrix data, then a feature selection is done by reducing the column from the matrix to parse the number of features, only features that are relevant to SVM-RFE and SVM-RPSO data are selected. Then the classification process is carried out on new data. Where classification in this research will compare the accuracy of SVM performed on SVM-RFE results data and SVM-RPSO results data. This is done to look for algorithms feature selection which is better accuracy in selecting features using the SVM-RFE method or SVM-RPSO method. SVM is used to use the Gaussian-RBF kernel function Kernel with σ = 0.05 and polynomial kernel with degree = 3.

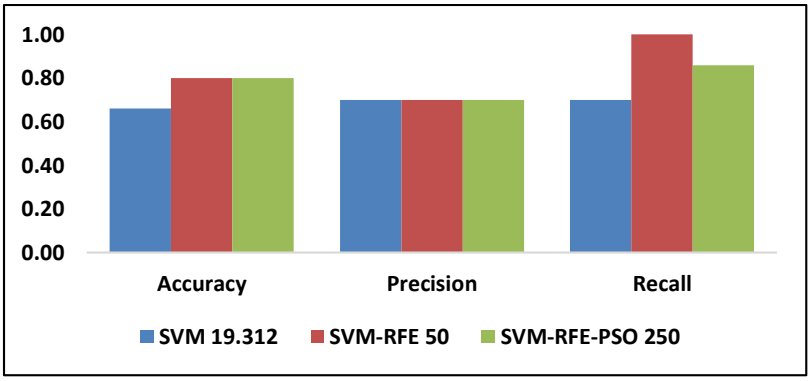

Figure 1. The performance comparison of SVM, SVM-RFE, and SVM-RFE-PSO methods using SVM linear.

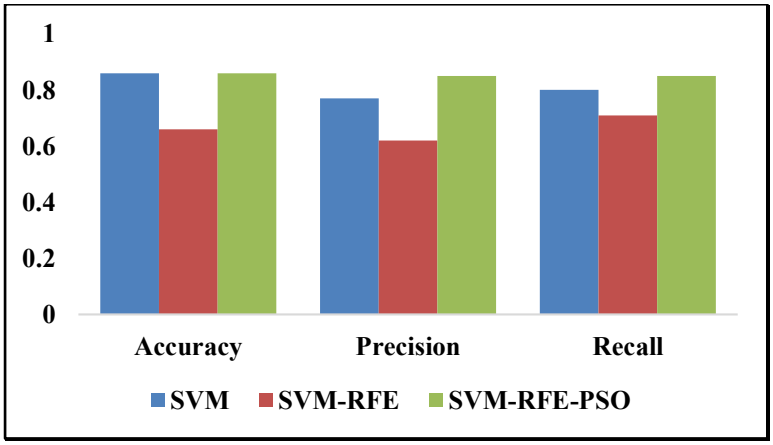

Figure 2. the performance comparison of SVM, SVM-RFE, and SVM-RFE-PSO methods using SVM polynomial kernel degree = 3.

Table 4. Accuracy % of classification (SVM Kernel Linear).0

Measure	SVM	SVM-RFE	SVM-RFE-PSO
Genes	19.312	50	250
Accuracy	66	80	80
Precision	70	70	70
Recall	70	100	85.71

Table 5. Accuracy of Classification (SVM Kernel Polynomial Degree = 3).

Measure	SVM	SVM-RFE	SVM-RFE-PSO
Genes	19.312	50	250
Accuracy	86.67 %	66.67 %	86.67 %
Precision	77.78 %	62.5 %	85.71 %
Recall	80 %	71.42 %	85.71 %

As the result based on the Tables 4 and 5 and Figures 1 and 2 we can see that the accuracy, precision and recall of the dataset have classified showed that SVM-RFE with optimization PSO has a better performance than using SVM as classifier and selection features based SVM-RFE.

Classification using SVM by using a polynomial kernel with a degree = 3 as a whole looks better than linear SVM in this data usage. The selection of features in the classification using the polynomial kernel SVM above shows that between SVM-RFE and SVM RFE PSO or in this study we call RPSO sheamed SVM has a higher standard, with accuracy, precision and recall, which is better with the use of the best features based on PSO with 250 features.

4. Conclusion

The current gene data microarray is a tool that can profiling gene expression and has been proven to be a value that determines the classification of complex diseases such as cancer and chronic diseases. But the progress of the development of knowledge about genes at this time makes information even bigger, or data becomes high dimension where in processing data becomes time consuming, memory, and cost. Therefore, it is necessary to have a method that can make the data processing process faster, namely the use of machine learning with optimization in the feature selection stage. Where by using an optimization process using PSO particle swarm optimization with SVM - RFE is expected to be able to choose the

optimal feature that can classify gene expressions from CKD data on GEO with 159 samples divided into 48 training and testing data samples 22 control samples and 26 CKD samples. 111 sample as validation data, 13 controls and 98 CKD samples.

As a result, it is seen in Tables 4 and 5 that the average performance of the data set processed with RPSO-SVM is better than just using SVM without feature selection stages with 81.48 %, RPSO-SVM is also better than SVM-RFE and then classified using SVM with 86.03 %. Overall the process of classification using SVM with 3 degree polynomial kernels and linear is better by using linear. Therefore, it can be concluded that the processing of CKD GSE97709 dataset microarray on NCBI By using RPSO-SVM and classification using kernel-based SVM better than SVM-RFE with ordinary RFE and SVM better with accuracy results higher than the others.

References

[1] Z.R.Yang, Mechine Learning Approaches to Bioinformatics, United Kingdom, (2009)
[2] M. Jhamb, K. Abdel, J. Yabes, Y .Wang, S. D. Weisbord, M .Unruh, J L Steel. Comparison of fatigue, pain and depression in patients with advanced kidney disease and cancer – symptom burden and clusters. *Journal of Pain and Symptom Management* Volume **57**, Issue 3, pp.566–575.e3 (2019)
[3] Y. Liu, J. Li, e J. Yu, Y. Wang, J. Lu, E.X Shang, Z. Zhu, J. Guo, J. Duan.Disorder of gut amino acids metabolism during CKD progression isrelated with gut microbiota dysbiosis and metagenome change. *Journal of Pharmaceutical and Biomedical Analysis* **149**, pp.425–435, (2018)
[4] P. Shi, S. Ray, Q. Zhu, M. A. Kon. Top Scoring pairs for feature selction in machine learning and applications to cancer outcome prediction. *BMC Bioinformaatics* **12** pp.3-15, (2011)
[5] Z. Rustam and N. Maghfirah, AIP Conference Proceedings 2023, 020235, (2018)
[6] Ressom H W, Varghese R_S, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. From Bioset **13**: 691-708. (2008)
[7] Y.Zhang, Q. Deng, W. Liang and X. Zou, An Efficient Feature Selection Strategy Based on Multiple Support Vektor Machine Technology with Gene Expression Data, *BioMed Research International volume 2018 7538204*, (2018)
[8] Kavitha, K., Harishankar, U. N., and Akhil, M. C, "PSO based feature selection of gene for cancer classification using SVM-RFE". in International Conference on Advances in Computing, Communications and Informatics (ICACCI). (2018).
[9] R.Indraswari, RBF Kernel Optimization Method with Particle Swarm Optimization on SVM using the analysis of input data's movement. Jurnal Ilmu Komputer dan Informasi (*Journal of Computer Science and Information*), Volume **10**, Issue 1, (2017)
[10] S.M. Ayyad, A.L. Saleh, L.M. Labib. Gene expression cancer classsification using modified K-Nearest Neighbords technique. *Journal BioSystems*, **176** pp.41-51, (2019).