A Hybrid Method for Multivariate Time Series Feature Selection

Ani Dijah Rahajoe[1] Edi Winarko[2] Suryo Guritno[3]
[1]Bhayangkara University, Surabaya, Indonesia,
[2,3]Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia.

Summary

Multivariate time series is usually used in multimedia, finance, medical, gesture and speech recognition. MTS dataset consists of m row and n column. Thus, classification or clustering would have a large size or a high dimensional space. The purpose of feature selection is to reduce the dimensionality without removing any information from the original variable. This paper contributes by means of new feature selection method based on observation times on each of its feature or variable. The proposed filter method uses the resulting factor loadings analysis from principle component analysis (PCA). The idea is to select features based on the time of observation that have most influences on other observation. Only observation times with the highest loading factor value from each principal component are counted. This method is called Feature Selection based Loadings factor (FSBLF) while the GASVM is used as the wrapper method. Fewer observations are included to make the running time faster than when using all observations. This means that only the selected features and observation times are used to perceive grade prediction in testing data. Not all features and observation times are used, and the original (raw/unprocessed) data are used. This method was compared between methods such as without feature selection (classification with SVM), GABayes, FSBLF, FSBLF_GABayes and FSBLF_GASVM. The proposed method has been tested to the CMU dataset and Wafer dataset. Results have shown the selection of a hybrid method FSBLF_GASVM has a few features with higher accuracy and only using the original data until the end of the feature selection process.

Key words:
Loadings, SVM, genetic algorithm, wrapper, filter, MTS.

1. Introduction

Feature selection has been used as a preprocessing technique before data mining process, such as classification or clustering, is performed. There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance [1]. When the data have abundant features, it will be difficult to identify and extract patterns or rules due to the high interdependence among individual features or the behavior of combined feature [2]. Feature selection tries to take relevant features and reduce the amount of irrelevancy or redundancy [3].

In general, feature selection algorithms can be categorized as[4]: filter, wrapper and embedded methods,. The filter method is an independent pre-processing step, independently of the classification. The goal is to find features that have qualities which may affect the outcome of the training without seeing the training mechanism selection . Essentially, irrelevant features are pre-filtered and then the generated feature subset is used for the training. Wrapper method uses machine learning as a black box to score subsets of features based on their predictive power[5,6].Wrapper method approach is often used in relation to learning or data mining algorithms, in contrast to the filter method and is part of the validation process. In the embedded method, feature selection in training process is usually more specifically used for learning machines. In this method, an explicit or implicit sub-algorithm feature selection is an integral part of a more general learning algorithm [7].
As we know, selecting features on MTS dataset is not easy. Each MTS represents an m x n matrix where m is the number of variables and n is the number of observations. The value of n is the number of observations in which t1,t2,...,tn represents time feature, while its variable is denoted as V1,V2,...,Vm. MTS uses the term variable to represent feature.. MTS data used in EEG [8,9], for Brain Computer Interface (BCI) are a representation of the channel used on the appliance [8,9,10]. Thus, in this research, feature, variable, and channel are used interchangeably. MTS dataset involves a lot of features and includes high-dimensional features. Some features on MTS data could be irrelevant and redundant due to the high number of features. It could reduce the accuracy of the classification. High-dimensional features also affect the pattern or image recognition system. In other words a lot of features reduce the accuracy of the result of recognition system classification which has the same non-informative features [11,12]. This is why the feature selection becomes the background of this research, to determine the level of accuracy of each algorithm. By reducing the number of features, classification models will become simpler and easier to understand.

In this research, feature selection is done through two main stages. The first stage uses filter method and the second stage uses wrapper method. The wrapper method is computationally costly, and it produces a good performance [13,2,14,15,6]. High processing cost is

---

Manuscript received March 5, 2017
Manuscript revised March 20, 2017

mitigated by the term hybrid to describe the combination of filter and wrapper methods whose objective is to reduce the computation time and improve the accuracy of classification result [16,17]. For this reason, the research is also based on hybrid filter wrapper.

This feature selection process selects any subset of features on a single MTS dataset. To be able to perform the process, each MTS dataset is grouped according to their variables. This study selects a subset of observation of each variable with a filter method followed by the wrapper method. Furthermore, the study sees which features are selected based on the ranking.

It is in contrast to the research [1,9] in which the feature subset selection techniques transform each MTS dataset into a single row and column called vectorization and use the Recursive Feature Elimination (RFE) on EEG dataset. Vectorization means to process data from multiple matrices into a single matrix. This research keep involves correlation through the principle components and the processes will be continued only by using the original data. This reasearch simply chose the most influential observation time in any channels or features. Each MTS Data Matrix is grouped according to their variables. In [9] have used the autoregressive model of order 3 for each channel without seeing the correlation between features, while [18] provides an overview of the preprocessing to transform an information data by looking at the correlation between variables.

The first feature subset selection is based on the loadings on the principle forming component. This research takes the value loading on the formation of the main components without multiplying it by the original data. PCA technique is taken to separate features that are dependent on other features. Thus, only variables that are independent were taken to the next process that we call Feature Selection Based Loadings Factor (FSBLF). The loadings value is also used in research [18] which contributes to the Descriptive Common Principal Component (DCPC). However, this research uses the result of DCPC for its vectorization, whose rows are its first PCs and columns representing variables. This is also true for [19] that forms its vectorization by multiplying the number of MTS by the minimum number of principal components and the number of original feature. In contrast, this research only selects variables with the greatest loadings of each principle component. The idea is to separate groups of feature subset in which each feature subset group has a strong correlation with one another. This research only takes the highest loading values that are the result of the first feature selection process. Thus, the selection results of the first subset of features is not dependent on other features. The results of the first feature selection stage are used as the input into the second feature selection stage, which uses GA-SVM wrapper method to find out which variables are selected through the rankings.

The remainder of this paper is organized as follows: In section 2, the related work is discussed. Section 3 provides our proposed method. Section 4 discusses the results and experiment. Conclusions are presented in Section 5.

2. Related Works

Feature subset selection methods such as Recursive Feature Elimination (RFE), Principle Component Analysis (PCA) and Support Vector Machine (SVM) are used to convert multiple dimensions of data into smaller dimensions. RFE in utilizes a classifier into the feature selection procedure to select the best features that produce the best classification precision [18]. [20] also claims that SVM, as a channel selection method, is a wrapper approach which recursively eliminates the least contributing channels based on the classification accuracy from SVM classifier. PCA is more widely used in dimensional reduction, which is commonly performed in the preprocessing stage as used in pattern recognition and classification applications [9].

Feature selection proposed by [18] is a cluster of unsupervised method called CLeVer (Descriptive Common principal Loading Component Based Variable Subset Selection Method) based on Common Principal Component Analysis method for FSS data from MTS. CLeVer is different from other FSS methods such as the Recursive Feature Elimination (RFE) and Fisher Criterion (FC) because these methods lose the information on the correlation between variables, while CLeVer does not lose information. This is because the CLeVer utilizes the properties of PCA techniques to maintain the correlation between variables.

The research in [21] uses CleVer as a comparison of the proposed methods, namely Class Separability Feature Selection (CSFS) for MTS feature selection by finding classseparation criteria. Feature selection algorithm (CSFS) selects features based on the score or class value and variable independently. MI or mutual information matrix between the variables is used as features for classification and is computed in every two variables. In general, MI uses criteria for feature selection [22]. MI is used for feature extraction because it can handle linear and non-linear data relation. Compared with CleVer, Corona and AGV, CSFS generally has better performance in four EEG data sets used in the research [21].

In [15] uses the wrapper method with GA to reduce the computational cost to find the optimal set of weights on a 38 dimensional feature set. The first GA phase is used to find the weight that minimizes speaker recognition validation error. The set of feature weight are calculated by averaging the optimal weight at 10 speaker subsets into two experiments. Then its feature selection is designed based on the feature ranking average .

Hybrid GA / SVM approach by [17] uses fuzzy logic to reduce the dimensions of the initial problem by eliminating the redundant genes. The next step selects genes with GA / SVM. GA is to find a subset of good genes, which are then evaluated by SVM. The dataset used is a dataset of patients with leukemia and colon cancer.

3. Proposed Method

This section will explain the proposed methods to be used for feature selection. As described earlier, the data MTS has m x n matrix dimensions. In this research, the classification process is simplified through feature selection in each time series observation. The goal of each time series observation is to find out the time that has strong influence on other features or variables. The impact will be proportional with the accuracy of the class prediction classification. The simplification of classification process through feature selection of hybrid method involves five processes that view correlations in each timeseries of the MTS data set.

The first phase of the process is the separation of each time series group based on the feature or variable or channel (e.g EEG). The separation of each time series group can be seen in Fig 1. For example, in the EEG data [23], some data of the patients have a time series of data from multiple channels (hereinafter referred as features or variables). The time series variable contains data of all patients.

The second phase of the process is done through uniformity of data dimension of each observation time, and it only involves additional data with 0 value for data with smaller dimension. This is done in order that PCA formation requiring similar dimension can be initiated. This is a coincidence that dataset have different lengths of observation times. If the used dataset have similar lengths of observation times, this phase is unnecessary.

In the third phase, feature selection process with the filter method is done by adopting PCA to see the loading value of any feature. PCA has been widely used for the analysis of multivariate data and dimension reduction [24]. The purpose of PCA is to find, select, and group the dominant variables affecting other variables so that the complex variables can be reduced without losing the information from the original variable. n variable will be reduced to a new variable m where m is smaller in number than n and eliminate the formed correlation [25]. PCA transforms variable x into a new variable y. The new variables are called the principal components. Generally, PCA steps start by calculating the average of each row in the variable and data matrices and reducing all values in each line with the average (Zero-mean). PCA is performed by applying Singular Value Decomposition (SVD) to either a matrix or a correlation matrix of an input data matrix [19]. Let A be a mean-centered MTS data of $m \times n$ dimensions and AA^T be roughly its $m \times m$ sample covariance matrix. Then, SVD decomposes the real, symmetric matrix AA^T is by Eq. (1):

$$AA^T = U \Lambda U^T \qquad (1)$$

where the columns of orthonormal matrix U are the m number of principal components and a diagonal matrix Λ has the corresponding variances along the diagonal [19]. Let the original variables be denoted by $x_1, x_2, ..., x_m$, then a principal component takes the form $u = \sum_i^m l_i x_i$, where l_i (i = 1,2,...,m) are often called as the *principal component loadings* and can be defined as the contributions or weights of original features loaded on determining the principal directions [19].

The formation of the principal components generate the loading value sequences of the greatest. Only original features loaded of the largest loading value of each principal component is taken. As a result, each matrix generates time observation that has been selected. We call it *Feature Selection Based Loadings Factor* (FSBLF). The loadings value is also used in Yoon et al. [18] which contributes to the Descriptive Common Principal Component (DCPC). The research used DCPC result for vectorization whose rows are its first p PCs and columns represent variables where p is number of PCs. Yoo et al. [19] established vektorization from number of MTS multiplied minimum number of principal components and number of original feature. In contrast, this research only selects variables with the greatest loadings of each principle component. The idea is to separate the groups of feature subset that have high relevance or correlation. The result is an observation feature that is no longer dependent on other feature, which is then used for training set and testing set.

The fourth phase of the process is wrapper method. The use of this method through GA-SVM is used to search for features of the selected times, and it has the highest accuracy for classification. GA is used to randomly select a feature which will be evaluated using the SVM. Genetic Algorithm - Support Vector Machine (GA-SVM) is one of the feature selection algorithms in the wrapper method category. Genetic algorithms belong to a class of population-based stochastic search algorithms that are inspired from principles of natural evolution known as evolutionary algorithms (EA)[26,25,27]. GA was designed to optimally solve sequential decision processes more than to perform function optimization but over the years, it has been used widely in both learning and optimization problems [25, 28]. GA has several stages that are: Encoding of chromosomes, initialization of the population, fitness function, selection of parents, crossover, mutation and population replacement, criteria restrictions. Genetic strings of very fit individuals are likely to influence the strings contained in the next generation whereas the strings of individuals with lower fitness are not likely to do [8]. There are several steps in the GA-SVM application, which are:

1. There are training and testing sets.
2. In the training set, features are selected randomly using GA.
3. The selected feature's accuracy is tested using SVM and data used is the testing set.
4. The steps are repeated with certain restrictions to obtain the highest level of accuracy and selected features.

GA intended for feature selection has different ways with those of GA for optimization. GA for feature selection has a purpose to select the features that will be used for the classification process with SVM. Thus chromosomes in GA are a set of selected and non-selected features. The population in one generation is a collection of some chromosomes that are comprised of selected or non-selected features, which are then tested with SVM to determine the accuracy level. The process of SVM is in every GA interaction process in each generation.

The fitness function of an individual is determined by evaluating the SVM. In this paper, we use one criterion fitness function containing accuracy for testing dataset as mentioned in Eq. (3):

$$fitness\ (x) = accuracy(x) \tag{3}$$

Accuracy (*x*) is the accuracy of the SVM classifier trained using the feature subset of training data represented by *x*.

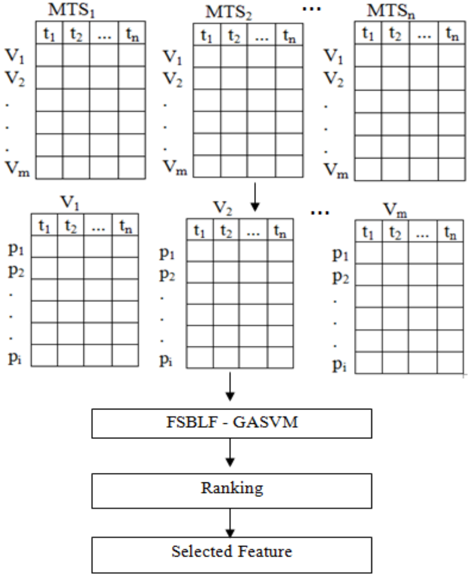
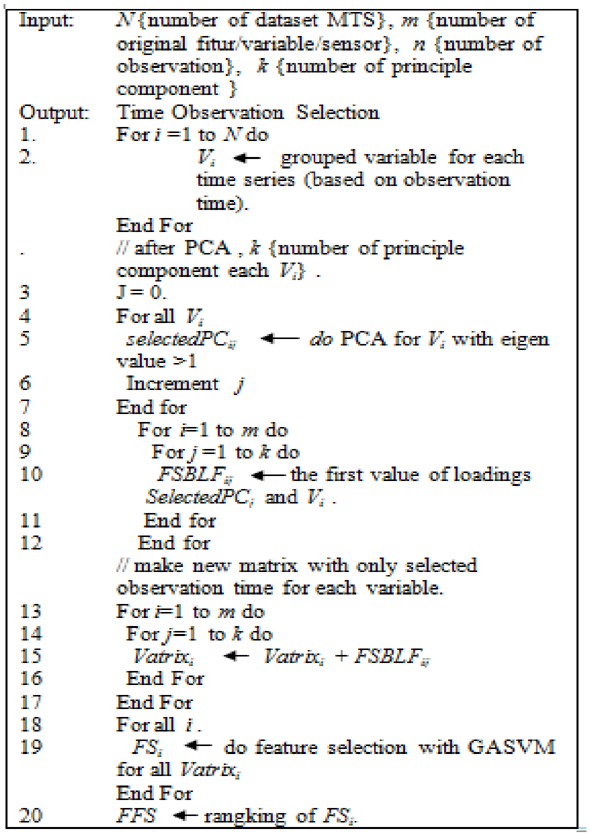


Fig. 1. General overview of proposal approach.



```
Input:   $N$ {number of dataset MTS}, $m$ {number of
         original fitur/variable/sensor}, $n$ {number of
         observation}, $k$ {number of principle
         component }
Output:  Time Observation Selection
         For $i$ =1 to $N$ do
                 $V_i$ ← grouped variable for each
                 time series (based on observation
                 time).
         End For
         // after PCA , $k$ {number of principle
         component each $V_i$} .
         J = 0.
         For all $V_i$
          $selectedPC_{ij}$ ← do PCA for $V_i$ with eigen
         value >1
          Increment $j$
         End for
            For $i$=1 to $m$ do
              For $j$ =1 to $k$ do
                $FSBLF_{ij}$ ← the first value of loadings
               $SelectedPC_j$ and $V_i$ .
               End for
            End for
         // make new matrix with only selected
         observation time for each variable.
         For $i$=1 to $m$ do
          For $j$=1 to $k$ do
           $Vatrix_i$ ← $Vatrix_i$ + $FSBLF_{ij}$
          End For
         End For
         For all $i$ .
           $FS_i$ ← do feature selection with GASVM
         for all $Vatrix_i$
         End For
         $FFS$ ← rangking of $FS_i$.
```

Fiq. 2. Feature Selection of FSBLF-GASVM Algorithm

MTS dataset used in this research consisted of a training dataset and testing dataset. Therefore, we use testing dataset to test its accuracy by using SVM. In conclusion, we select or screen observation times for every variable or feature from dataset. The first screening involves filtering method by using FSBLF, and the second is done with GA-SVM. Vectorization is done by feature or variable. This is a new method from the perspective of the utilized variable(s). The most important issue in this research is that it is uses the original data through the processes. The next question is then how to select the screened features or variables.

This fifth phase is based on every variable or feature that has the highest classification accuracy with the shortest observation times that have been selected. Therefore, through the selection there will be a small number of variables with certain times and the highest accuracy. With this selection predicting grades in the testing data does not require all of the data (features and observation times). This means that if there are two selected variables or channels and each of the three observation times at *t*, only these features and observation times will be used to perceive the prediction grade in the testing data.

The general overview of this research can be seen in Fig 1. After selection of observation time, each subsequent variable make the selection of observation time with the wrapper method. The ranking of the selected feature is based on the results of the highest accuracy with the fewest selected feature subset. Fig. 2 shows the steps of feature selection of the FSBLF-GASVM algorithm and shows that each feature selection process still uses the original data for each variable. It is different with most researches, which transform variables into a column or a row with extraction data [19]. This research still uses the observed values selected for the final result.

4. Performance Evaluation

This study uses parameter to apply GA-SVM. The parameter can be seen in table 1. The selection strategy used in this study is Roulette Wheel [29]. We assess one point crossover and generational replacement. For the classification of SVM the researchers used kernel linear function.

Table 1. Parameter of performance evaluation.

GA Parameter	**Value**
Size of population	50
Number of maximum generation	100
Crossover rate	0.8
Mutation rate	0.05

4.1 CMU and Wafer Dataset

CMU Motion Capture is used to see the performance of the proposed method [30]. MTS dataset is the data for the motion or gesture of a person (subject 16) based on command. This dataset has 2 classes. The commands or tasks are to predict whether the subject is walking or running. This dataset has information from 62 different combined positions recorded [31]. These positions represent data variable that can be denoted by V1, V2,...Vn. Some data features (variable) have the same value. These variables are variables 34 and 46. Both variables cannot be observed because it can not find the value of loading. In addition, there are several variables that cannot be observed because their values are too small (less than o). These variables are variables 25, 26, 37, 38. Therefore, these variables are not included in the research.

Wafer dataset is discussed by Olszewski [32, 33]. The wafer database comprises a collection of time-series data sets where each file contains the sequence of measurements recorded by one vacuum-chamber sensor during the etch process applied to one silicon wafer during the manufacture of semiconductor microelectronics. Each wafer has an assigned classification of normal or abnormal. The abnormal wafers are representative of a range of problems commonly encountered during semiconductor manufacturing. Table 2 shows summary of dataset structure.

Table 2. Dataset Structure

Dataset	# of variabels	Length	# of instances	
			Training	Testing
CMU_S16_ MOCAP	62	127-580	29	29
Wafer	6	104-198	298	896

4.2 Performance of CMU Dataset Classification.

Datasets are grouped based on the variable. Each variable contains a collection of data subject and the time of observation. Thus, there will be 62 new matrices. Each matrix has 29 time series. CMU_Subject16 has the length of time series between 127-580 so that each matrix has a dimension that ranges from 29x127 to 29x580. The difference of the time series length is unified with a value of 0 in return. After unifying the dimensional matrix, the next phase will produce 62 matrices. Each matrix has a dimension of 29x580. Each dimension is used as the data in this phase.

The implementation of FSBLF is compared in several methods. The goal is to determine its effectiveness and performance. It is effective because the selected features are fewer and they have a high degree of accuracy. The first step is to see the value of accuracy without feature selection, but only with SVM classification. The second step it to use the wrapper - GABayes. It uses genetic algorithm as a method of searching and Naïve Bayes for evaluators and classification. The third step is to compare it with the hybrid filter wrapper (GASVM, GABayes). Especially for GABayes in this research using Weka app version 3.6.11.

Fig. 3 shows a comparison of the accuracy with 5 different methods. FSBLF-GASVM has 3 variables with 100 percent accuracy with at least two selected features. The first feature is the variable 3 at t93 and t235, the second feature is the variable 40 at t1 and t4.

Judging from the stability graph, the variables selected from FSBLF_GASVM are more accurate than those from other methods. The accuracy of GABayes method is close to that of FSBLF_GASVM but the number of selected features is higher than that of FSBLF_GASVM. Thus, GABayes is less effective. The number of selected features can be seen in Fig. 4.

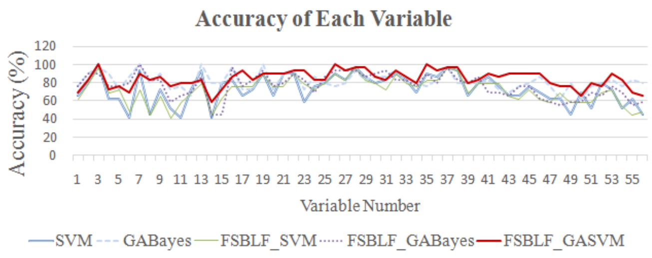


Fig. 3. Accuracy of five methods for CMU MOCAP S16.

In Fig.4, the number of selected features are in hundreds. Therefore, the computational time is still high or it takes longer than hybrid filter wrapper method.

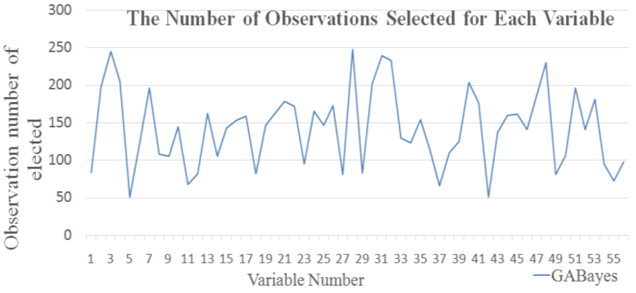


Fig. 4. The number of selected time features obtained by using only GA-Bayes wrapper methods on CMU dataset.

Fig.5 shows the number of variables or observation feature. Although the FSBLF_GABayes method has fewer selected time feature, but the accuracy values in Fig. 4 are still better than those of the FSBLF_GASVM method. FSBLF_GABayes has a slightly lower accuracy level compared with that of FSBLF_GASVM. If we only use FSBLF, the accuracy level is still much smaller than that of FSBLF_GASVM. In FSBLF_GASVM, two variables gain 100 percent accuracy. Each of these variables generates 2 selected observation time. These variables are variables 3 and 40. This happened to variables 3 at t93 and t235, variable 40 at t1 and t4. The variable 28 has seven observation time selection with 100 percent accuracy rate.

The last phase is to see which features are selected. The details of ranking acquisition of each method can be seen in Fig. 6. From this figure, FSBLF_GASVM hybrid method obtains the maximum accuracy with the least amount of selected observation time. FS is the number of selected features of each method used on CMU data. Obs represent the number of selected observation time. FSBLF_GASVM method has fewer feature results with maximum accuracy. Thus the proposed method is more accurate than when using these other methods: SVM classification without feature selection, GABayes wrapper method only, FSBLF only (FSBLF results, classified by using SVM), a hybrid method of FSBLF_GABayes and FSBLF_GASVM. To find out more about FSBLF performance, see Fig. 7.

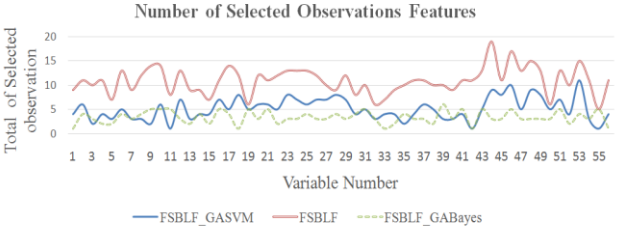


Fig. 5. Number of observations selected from each variable on CMU dataset.

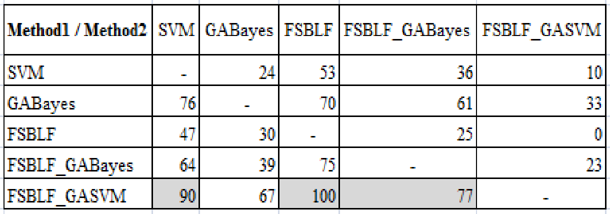
In Fig. 7, the accuracy of FSBLF_GASVM is 100 percent higher than that of FSBLF. The value indicates a percentage of accuracy result of the first method to the second method. This means that each classification result of each variable in FSBLF_GASVM has higher accuracy than those in FSBLF. The classification of FSBLF_GASVM is 90 percent greater than that without feature selection (SVM). This means that only 10 percent has lower accuracy. If FSBLF_GASVM is compared with FSBLF_GABayes, the accuracy rate of FSBLF_GASVM is higher by 77 percent. The proposed method (FSBLF_GASVM) has a better performance level than those of FSBLF_GABayes, SVM, GABayes and FSBLF.

4.3 Performance of Wafer Dataset Classification

The results of the proposed method are compared to those of some other methods. Those methods are classification without selection (direct classification with SVM), GABayes feature selection method, and filter wrapper method or hybrid FSBLF_GABayes and FSBLF_GASVM. Fig. 8 shows that the proposed method (FSBLF_GASVM) has a better accuracy rate than the four other methods. We use GA-SVM because studies [33] state that GASVM has better performance than PSOSVM. For comparison, we use GABayes which has the same genetic algorithm base. Accuracy results can be seen in Fig. 8.

Rangking	GABayes			FSBLF_SVM			FSBLF_GABayes			FSBLF_GASVM		
	FS	Obs	%	FS	Obs	%	FS	Obs	%	FS	Obs	%
1	V19	147	100	V3	9	96,5	V7	3	100	V3	2	100
2	V13	163	100	V30	10	96,5	V42	3	96,5	V40	2	100
3	V7	196	100	V42	11	96,5	V16	5	96,5	V28	7	100
4	V42	66	96,5	V43	10	93,1	V28	3	93,1	V42	6	96,5
5	V16	153	96,5	V19	6	89,6	V29	3	93,1	V31	7	96,5

Fig. 6. Number of observation results and accuracy from four methods on the CMU dataset.



Method1 / Method2	SVM	GABayes	FSBLF	FSBLF_GABayes	FSBLF_GASVM
SVM	-	24	53	36	10
GABayes	76	-	70	61	33
FSBLF	47	30	-	25	0
FSBLF_GABayes	64	39	75	-	23
FSBLF_GASVM	90	67	100	77	-

Fig. 7. The first and second comparisons of accuracy level from each method from CMU dataset.

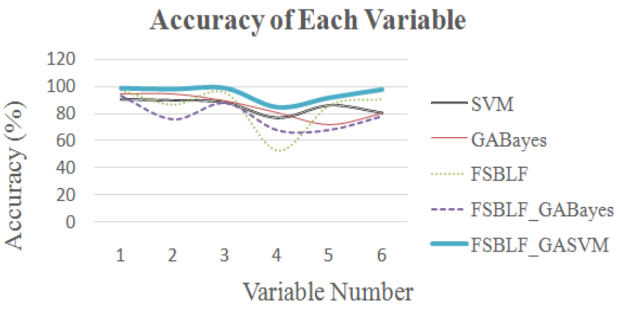


Fig. 8. Accuracy of five feature selection methods.

Fig. 9 shows the amount of feature selection obtained from the five methods used. FSBLF-GASVM has a higher number of feature selection methods than that of FSBLF_GABayes, but FSBLF_GASVM still has a higher accuracy.

The ranking features can be seen in Fig. 10. FS is the number of selected features of each method used on CMU data. Obs represent the number of selected observation time. This image shows the proposed method (FSBLF_GASVM) has a better accuracy rate than the other three methods.

In Fig. 11, the method in the list of line is compared with the method in a vertical list or column. The value indicates a percentage of accuracy result of the first method to the second method. FSBLF_GASVM had a 100 percent higher performance than SVM, GABAyes, FSBLF, FSBLF_GABayes. Therefore, the performance of FSBLF_GASVM is better than the four other methods.

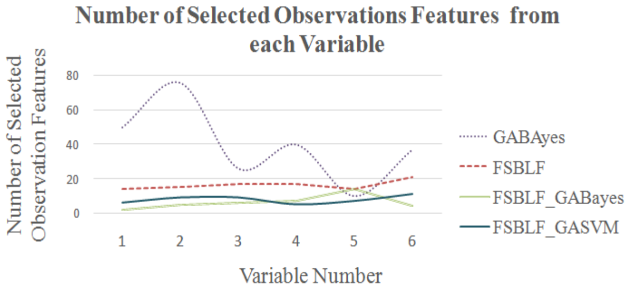
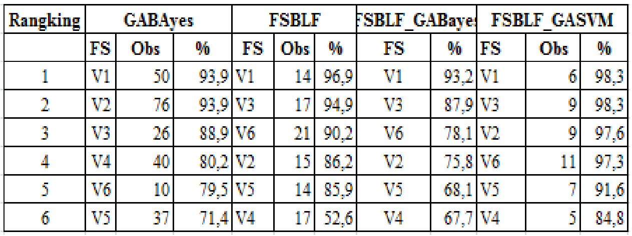
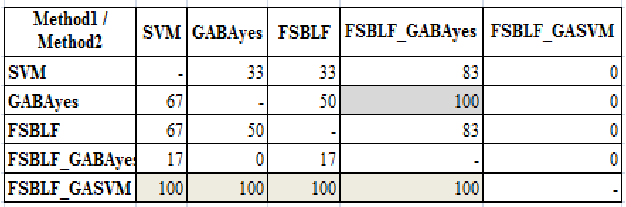


Fig. 9. Number of observation results from the four methods on the wafer dataset.



Rangking	GABAyes			FSBLF			FSBLF_GABaye		FSBLF_GASVM		
	FS	Obs	%	FS	Obs	%	FS	%	FS	Obs	%
1	V1	50	93,9	V1	14	96,9	V1	93,2	V1	6	98,3
2	V2	76	93,9	V3	17	94,9	V3	87,9	V3	9	98,3
3	V3	26	88,9	V6	21	90,2	V6	78,1	V2	9	97,6
4	V4	40	80,2	V2	15	86,2	V2	75,8	V6	11	97,3
5	V6	10	79,5	V5	14	85,9	V5	68,1	V5	7	91,6
6	V5	37	71,4	V4	17	52,6	V4	67,7	V4	5	84,8

Fig. 10. Number of observation results and accuracy from four methods on the wafer dataset.



Method1 / Method2	SVM	GABAyes	FSBLF	FSBLF_GABAyes	FSBLF_GASVM
SVM	-	33	33	83	0
GABAyes	67	-	50	100	0
FSBLF	67	50	-	83	0
FSBLF_GABAye	17	0	17	-	0
FSBLF_GASVM	100	100	100	100	-

Fig. 11. The first method compared with the second method form Wafer dataset.

5. Conclusion and Future Work

This research still uses the original data without feature extraction. For each time series, the time observations take only the most influential and highest classification accuracy values. The results shows that the proposed method of feature selection (FSBLF) can be used for time series observation feature selection. FSBLF-GASVM has a better performance than the four other methods (SVM, GABayes, only FSBLF, FSBLF_GABayes). The results of the selected feature in the first stage are much fewer. Therefore, processing will be faster and more varied. Although GABayes wrapper method has better accuracy, the number of feature selection result is over a hundred. FSBLF-GASVM results on CMU data has 100 percent accuracy rate in two-time observation in one variable. This happened to variables 3 at t93 and t235, variable 40 at t1 and t4 and variable 28 with seven time observation. This means that it is adequate that we choose one of the selected variables and its observation times to see the grade prediction in the testing data. It is unnecessary to use all of the data variables and observation times in training dataset. In addition, it is unnecessary either to process or to extract data from the raw (initial) data to see the grade prediction from the data of a new patient.

The Wafer dataset on variable 1 has 6 selected observation time features and variable 3 has 9 selected observation time features . The result reaches the accuracy of 98.3

percent. Therefore, to see the grade prediction only requires variable 1 with six observation times.

A further research is needed to determine the performance of the classification of the proposed method by comparing it with other methods, especially with other soft computing wrapper methods.

Acknowledgments

The authors are grateful to the Universitas Gadjah Mada, Yogyakarta, Indonesia , for the full laboratorium support granted to this research.

References

[1] I. Guyon, "An Introduction to Variable and Feature Selection," vol. 3, pp. 1157–1182, 2003.
[2] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," in ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 601–608.
[3] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis, vol. 1, no. 1–4, pp. 131–156, 1997.
[4] H. Chouaib, O. R. Terrades, S. Tabbone, F. Cloppet, N. Vincent, and B. P. Nancy, "Feature selection combining genetic algorithm and Adaboost classifiers," vol. 5, no. Ea 2517, 2008.
[5] I. Oh, J. Lee, and B. Moon, "Hybrid Genetic Algorithms for Feature Selection," IEEE TRansaction on Pattern Analysis And Machine Intellegence, vol. 26, no. 11, pp. 1424–1437, 2004.
[6] R. Kohavi and H. John, "Artificial Intelligence Wrappers for feature subset selection," vol. 97, no. 97, pp. 273–324, 2011.
[7] I. . Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, 2005.
[8] M. Schroder, M. Bogdan, and W. Rosenstiel, "Automated EEG Feature Selection for Brain Computer Interfaces," pp. 1–4, 2003.
[9] T. N. Lal, S. Member, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf, "Support Vector Channel Selection in BCI," IEEE Trans. Biomed. Eng, vol. 51, no. 6, pp. 1003–1010, 2004.
[10] K. Yang, H. Yoon, and C. Shahabi, "A Supervised Feature Subset Selection Technique for Multivariate Time Series," in FSDM, 2005.
[11] B. Oluleye, A. Leisa, and D. Dean, "A Genetic Algorithm-Based Feature Selection," vol. 5, no. 4, pp. 899–905, 2014.
[12] L. Bruzzone and C. Persello, "A novel approach to the selection of robust and invariant features for classification of hyperspectral images," pp. 66–69, 2008.
[13] G. C. Centre, "A Genetic Algorithm Based Wrapper Feature Selection Method For Classification of Hyperspectral Images," pp. 397–402, 2006.
[14] E. P. Ephzibah, "Cost Effective Approach on Feature Selection Using Genetic Algorithms And Fuzzy Logic For Diabetes Diagnosis .," vol. 2, no. 1, pp. 1–10, 2011.
[15] M. Zamalloa, G. Bordel, L. J. Rodríguez, and M. Peñagarikano, "Feature Selection Based on Genetic Algorithms for Speaker Recognition," IEEE Speaker and Language Recognition Workshop, pp. 1–8, 2006.
[16] C. huei Yang and L. Chuang, "IG-GA : A Hybrid Filter / Wrapper Method for Feature Selection of Microarray Data," Medical and Biological Engineering, vol. 30, no. 1, pp. 23–28, 2009.
[17] E. B. Huerta, "A Hybrid GA / SVM Approach for Gene Selection and Classification of Microarray Data," pp. 34–44, 2006.
[18] H. Yoon, K. Yang, and C. Shahabi, "Feature Subset Selection and Feature Ranking for Multivariate Time Series," vol. 17, no. 9, pp. 1186–1198, 2005.
[19] H. Yoon and C. Shahabi, "Feature Subset Selection on Multivariate Time Series with Extremely Large Spatial Features ∗ ," vol. 0238560, 2006.
[20] M. Arvaneh, S. Member, C. Guan, S. Member, and K. K. Ang, "Optimizing the Channel Selection and Classification Accuracy in EEG-Based BCI," vol. 58, no. 6, pp. 1865–1873, 2011.
[21] M. Han and X. Liu, "Feature selection techniques with class separability for multivariate time series," Neurocomputing, vol. 110, pp. 29–34, Jun. 2013.
[22] G. Doquire and M. Verleysen, "Feature selection with missing data using mutual information estimators," Neurocomputing, vol. 90, pp. 3–11, Aug. 2012.
[23] I. Rejer and K. Lorenz, "Genetic algorithm and forward method for feature selection in EEG feature space," Journal of Theoretical and Applied Computer Science, vol. 7, no. 2, pp. 72–82, 2013.
[24] I. T. Jolliffe, Principal Component Analysis, Second Edition. New York: Springer-Verlag, 2002.
[25] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, "A novel ACO–GA hybrid algorithm for feature selection in protein function prediction," Expert Systems with Applications, vol. 36, no. 10, pp. 12086–12094, Dec. 2009.
[26] C. M and M. Z, "Evolutionary computation control and cyberneticse," in Proceedings of the IEEE , 26(3), 1997, pp. 307–338.
[27] M. E. Basiri and S. Nemati, "A Novel Hybrid ACO-GA Algorithm for Text Feature Selection," pp. 2561–2568, 2009.
[28] S. A and T. H, "A comparison between genetic algorithms and sequential quadratic programming in solving constrained optimization problems," ICGST International Journal on Artificial Intelligence and Machine Learning (AIML), vol. 6, no. 1, pp. 67–74, 2006.
[29] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Third edit. New York: Springer-Verlag, 1996.
[30] R. T. Olszewski, "'http://www.cs.cmu.edu/ebobski/,' accessed: May 9, 2016.".
[31] A. McGovern, D. H. Rosendahl, R. A. Brown, and K. K. Droegemeier, "Identifying predictive multi-dimensional time series motifs: An application to severe weather prediction," Data Mining and Knowledge Discovery, vol. 22, no. 1–2, pp. 232–258, 2011.
[32] R. T. Olszewski, "Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data," 2001.
[33] E. Alba, L. Jourdan, and E. Talbi, "Gene Selection in Cancer Classification using PSO / SVM and GA / SVM Hybrid

Algorithms," IEEE Congress on Evolutionary Computation (CEC), vol. 07, pp. 284–290, 2007.

Ani Dijah Rahajoe is currently a PhD student at the Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia. She received MCs. degree in Computer Science and Electronics from Universitas Gadjah Mada, Yogyakarta, Indonesia. Her research interests are in feature selection, data mining and soft computing. Email: anidrahayu@gmail.com.

Edi Winarko is currently a head of doctoral program in Computer Science, Department of Computer Science and Electronics at the Universitas Gadjah Mada in Yogyakarta, Indonesia. He holds a PhD degree in Computer Science School of Informatics and Engineering Flinders University, Australia. His research interests are in data mining, big data computing. Email: ewinarko@ugm.ac.id.

Suryo Guritno is an associate professor in Department of Computer Science and Electronics at Universitas Gadjah Mada, Yogyakarta, Indonesia. He received M.Stats. degree at The University of New South Wales, Sydney, Australia. He holds a Ph.D. degree at University of Connecticut, Storrs, USA. His reaseach interests are in statistic, computer science. Email: suryoguritno@ugm.ac.id.