ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse16862019.pdf
https://doi.org/10.30534/ijatcse/2019/16862019

A Wrapper Feature Selection Based on Ensemble Learning Algorithm for High Dimensional Data

Maryam[1], Noor Akhmad Setiawan[2]
[1] Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Indonesia
[2]Departmetn of Electrical and Information Engineering, Universitas Gadjah Mada, Indonesia

ABSTRACT

High dimensional data with limited amount may cause classification process more difficult. Certain technique, such as dimension reduction, is needed in order to simplify the classification model. Dimension reduction produces the best feature subset which makes classification process easier and gives improvement of result accuracy. This research applies wrapper method of Genetic Algorithm using ensemble learning AdaBoost as reduction algorithm. Algorithm Support Vector Machine is used as evaluator. Three datasets are taken from the UCI Machine Learning Repository. On Naïve Bayes classifier, accuracy rates of dermatology, heart, and primary tumor with GA-AdaBoost scheme are 88.19%, 58.65%, and 25.38% respectively. On k-NN classifier, the accuracy rates are 88.19%, 57.88%, and 37.69%. On Random Forest Tree classifier, the accuracy rates are 92.87%, 57.09%, and 31.21%. The proposed method gives statistically significant increase of accuracy and is superior compared to other feature selection schemes. This proves that the proposed GA-AdaBoost feature selection scheme is promising to handle high dimensional data.

Key words : Classification, high dimensional data, ensemble learning algorithm, dimension reduction.

1. INTRODUCTION

High dimensional data is often seen in real-world domain such as medical record data, knowledge data, and business data. The higher the data dimension, the amount of sample needed increases exponentially [1], hence the number of available data must be sufficient to avoid overfitting. Data mining process gets more difficult due to big data complexity. Phenomenon related to problems of high dimensional data is often called as “curse of dimensionality” (Bishop, 1995). The more amounts of dimensions, the higher amounts of data distributions. On classification process, if the number of observation data is small, there is no sufficient variable to generate reliable model for class labeling of all existing variable [2].

Performance enhancement on high dimensional data can be done by reducing dimensions. Such process makes data processing more effective and efficient. One way to reduce dimensions is feature selection. Feature selection can eliminate irrelevant features, reduce noises, and give better prediction results [3]. Feature selection algorithm may improve inductive learning in terms of generalization capacity, learning speed, and reducing model complexity as it is built using less features.

The proposed method describes wrapper feature selection algorithm, Genetic Algorithm, which will produce optimal feature subset. Ensemble learning method is used for partitioning data into different segments using AdaBoost method. Data sample of bootstrap training is taken from iteratively renewed dataset, hence the following classification can focus on difficult segments. Algorithm evaluator employed is Support Vector Machine (SVM). The combination of these methods is expected to be reliable in reducing dimensions and able to enhance performance on high dimensional data.

This research paper is divided into several sections. Section 2 describes previous literatures on wrapper feature selection and ensemble learning algorithm. Section 3 explains basic concept of the method employed. Section 4 illustrates design of the proposed method. Section 5 shows details of results and analysis. Section 6 describes conclusion of results of the study.

2. LITERATURE REVIEW

High dimensional data are likely to have irrelevant, excessive attributes and noise [4]. Common method used for reducing data dimensions is by reducing features. Feature selection forms optimal feature subset which is able to present equal or better performance compared to the original dataset. Another advantage is its ability to shorten training time of induction algorithm, reduce computing cost, and make data processing outcome easier to be discerned. Feature selection can be conducted using wrapper method: using induction algorithm as a part of feature selection. Since wrapper method collaborates with induction algorithm, this method is more likely to generate better performance compared to filter method [5].

Researches on feature selection for cases of high dimensional data that uses high dimensional data have been conducted. Principal Component Analysis (PCA) method is used as high

dimensional data feature selection and SVM is employed as classifier with RBF kernel optimization. As results, SVM classifier has higher efficiency than other classifier [6]. Researches on hybrid Genetic Algorithm (GA) method is done for improving feature selection [7]. Relief and Correlation based Featured Selection (CFS) are used as data input for GA wrapper. Hybrid technique from combination of feature selection methods of relief and GA, also Naïve Bayes as evaluator stated that proposed method gives fair outcomes.

GA is used as wrapper feature selection and SVM is used as algorithm evaluator for large biomedical dataset. The results of the research shows that GA and SVM are able to perform well and give accuracy improvement, using four binary class datasets and one multiclass dataset [8]. Combination of GA and SVM methods is tested with six binary class datasets [9]. The research compared SVM results with k- Nearest Neighbour (k-NN), Decision Tree, and Linear Discriminant Analysis. The outcomes showed that GA-SVM gave better accuracy results compared to combination of GA and other classifications. Hybrid chi square method and GA, alongside with SVM as evaluator are used as proposal of reliable feature selection method as results of researched showed the method proposed may significantly increase performances on some high dimensional data [10]. A new wrapper approach [11], using Incremental ANOVA and Functional Network to select feature used for dealing with classical algorithm with multiclass problems, such as C4.5 and Naïve Bayes. This method achieve better accuracy result.

Wrapper method may be optimized by using ensemble method. Reviews and comparison of methods using two techniques are feature selection and ensemble learning [12]. Ensemble learning feature selection gives higher accuracy than conventional methods of feature selection. Ensemble learning gives stability handling ability, therefore it is able to maximize the performance of feature selection [3] and for classification [13]. Ensemble learning is also used to reduce dimensions on high dimensional data [14]. Ensemble algorithm for classification is run by combining conventional feature selection, stating that the proposed method may increase its accuracy [15]. A new method used ensemble learning and swarm intelligent based feature selection for Cleveland heart disease prediction, the result of the experiment has been reduced by the fact that classification and prediction are considerably improved. From the accuracy perspective, those techniques provide an accuracy more than 95% compare with others [16].

Previously discussed literatures give opportunity to optimize the performance of GA wrapper on SVM therefore it works properly on problems of high dimensional data. One way to maximize feature selection performance is by using ensemble learning on wrapper method. Hence, a proposed method will be developed in this research. It is GA wrapper method based on AdaBoost ensemble learning technique using SVM as evaluator which is proven to work smoothly alongside with GA wrapper.

3. METHODOLOGY

The proposed method for this research is divided into three sections: datasets employed, feature selection technique, and evaluation.

3.1 Dataset

Three high dimensional biomedical datasets are used in this research. These datasets from the UCI Machine Learning Repository. Table 1 shows details on the datasets employed. Each has large number of features with numerous classes; hence it is classified as high dimensional data.

Table 1: Detail of dataset using in this study

No	Dataset	Data	Feature	Class
1	Dermatology	366	34	6
2	Heart	303	13	5
3	Primary	339	17	22

3.2 Feature Selection Technique

This research uses Genetic Algorithm based on AdaBoost ensemble learning and SVM classifier as evaluator. The result of feature selection is evaluated using classification algorithms to see the reliability of the proposed method in solving cases of high dimensional data. The datasets used have undergone data preprocessing such as replacement of missing value using mean of the data and normalization using Z-transformation procedure.

Dataset full feature is divided into three sections with the same size, which are data validation, data training, and data evaluation. Data validation is used for searching optimal parameter of c and from SVM using gridsearch and 10-fold cross validation method. Data training is used on GA wrapper and AdaBoost method. Data evaluation is used by taking best feature subset and evaluated using classifier algorithm. Figure 2 shows feature selection process using Genetic Algorithm wrapper and AdaBoost ensemble learning using SVM as evaluator.

A wrapper approach is used to calculate attribute weights. The inductive algorithm is used by wrapper methods as the evaluation function. After statistical resampling or cross-validation of the sample, this technique uses a classifier to assess subsets by their predictive accuracy (on test data). The wrapper method also achieves higher recognition levels than a filter approach because the former is geared to the unique interactions between the classifier and the dataset. In addition, wrappers have a mechanism to avoid overfitting, as predictive accuracy measures are typically used for cross-validation [3].

Genetic algorithm is a search algorithm based on natural selection and genetics mechanisms [7] which performance is determined by several parameters. These parameters are max number of generations, pop size value, probability mutation, and probability crossover. The first step in searching for GA

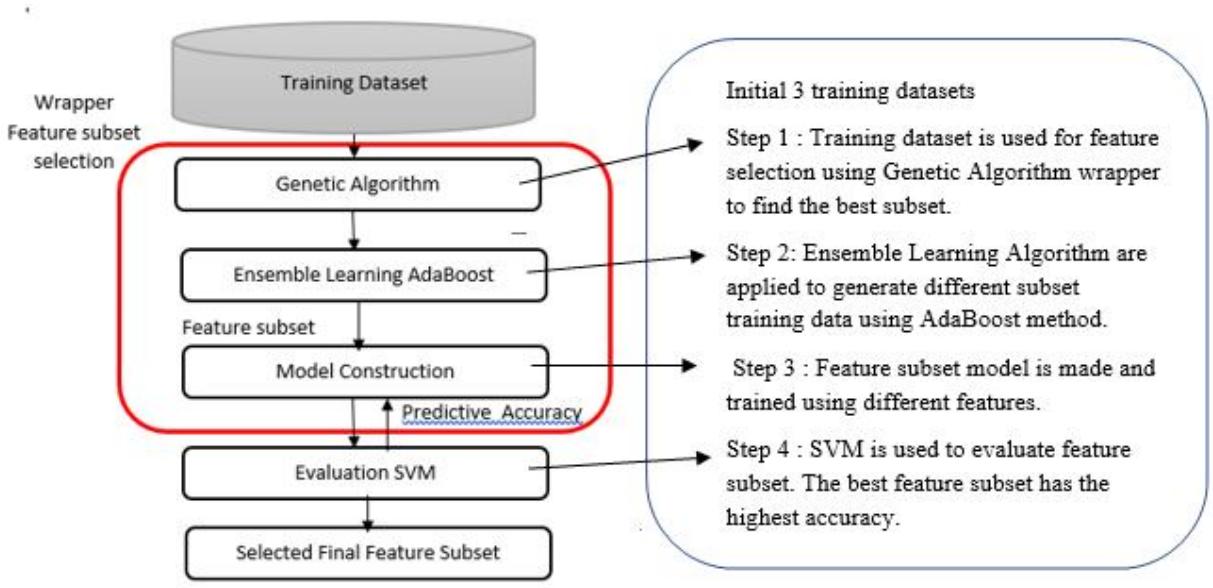

Figure 1: Genetic Algorithm based on Ensemble Learning Adaboost for feature selection

parameter is population initiation. The gene value of the chromosomes is scored, and chromosome with the highest fitness value becomes the strongest candidate for the next step. If it has not reached the maximal number of generations, the iteration will continue running. The chosen chromosome will be 'crossover' based on its crossover probability value. After that, the number of gene in the chromosome, which will be mutated, will be determined based on its probability mutation value. After reaching maximum generation, chromosome with highest fitness value will be obtained as best feature subset [8].

AdaBoost (Abbreviation for Adaptive Boosting) is meta algorithm which is able to work with other algorithm in order to improve performance. AdaBoost is proven empirically to improve the performance of generalization [17].

SVM implements "one-vs-all" method in dealing with high dimensional data with many classes. SVM is one of classifying algorithms which require settings in its kernel function. The recommended kernel function is the RBF kernel because it has the same performance as a linear kernel on certain parameters and has behavior with certain parameters like the sigmoid kernel function and a small range of values [0; 1] [9]. RBF kernel allows settings in parameter c and γ . SVM parameter optimization is done using grid search method, with range for each c and γ is $log_2 C \in \{-5, -3, \dots, 15\}$ and $log_2 C \in \{-15, -13, \dots, 3\}$.

3.3 Evaluation

Evaluation process is conducted using evaluation data based on selected subset feature using classification algorithm. Classification performance is done using 10-fold cross validation. This method, which divides datasets into 10-folds, is commonly used in data mining [18]. Classifiers employed in this research are Naïve Bayes, k-NN with k = 5, and RFT. These classifiers are selected due to its reliability in dealing with high dimensional data [19].

The compared performances are performances from datasets without feature selection process. Feature selection uses GA and GA AdaBoost combination. The evaluation is accuracy which is the probability of properly classified instances on dataset, as written in Equation 1.

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

According to (1), True Positive (TP) is the number of properly classified instances. False Negative (FN) is the number of instances from corresponding rows, except TP value. False Positive (FP) the number of instances from corresponding columns, except TP value. True Negative (TN) is the total of all columns and rows except the TP value assigned [20].

Accuracy is used for the significance test to see the significance of feature selection technique implementation. Significance test is run on three data groups: data without feature selection, data with GA feature selection, and data with GA-AdaBoost feature selection.

4. RESULTS AND DISCUSSION

4.1 Feature Selection

Feature selection is an important step in data mining implementation. Table 2 shows the number of features reduced by proposed method. The chosen feature subset is the best feature with highest accuracy by SVM. Feature selection method may reduce some features from the original dataset.

Figure 2 illustrates chart of SVM accuracy on each feature selection scheme. Generally, it can be inferred that the proposed feature selection combination, GA-AdaBoost, has better accuracy scores compared to when the features are individually implemented. In dermatology datasets, GA and GA-AdaBoost feature selection schemes reach equal accuracy rate, 99.33%. Meanwhile, heart and primary tumor datasets reach highest accuracy with GA-AdaBoost schemes as much as 68.46% and 46.92% respectively.

4.2 Classification Performance

Classification performance uses 10-fold cross validation and is average results of 30 performances resulted from data randomization. Performance matrix employed in this discussion is accuracy. Classifications used for evaluation comparison in each feature selection scheme are Naïve Bayes, k-NN with k=5, and RFT.

Accuracy rate is measured using significance test with significance rate of 5%. Table 3 shows classification performance resulted from Naïve Bayes, k-NN, and RFT classifier. Numbers written bold are performance results with the highest accuracy.

On Naïve Bayes application, GA-AdaBoost feature selection scheme may increase accuracy significantly on dermatology, heart, and primary tumor as much as 88.19%, 58.65%, and 25.38% respectively. On k-NN application, GA-AdaBoost feature selection scheme significantly decreases on dermatology dataset with 88.19% accuracy (lower compared to accuracy without feature selection), increases insignificantly on heart dataset with 57.88% accuracy rate,

Table 2: Number of features as resulted from feature selection scheme

No	Dataset	Full features	GA	GA-AdaBoost
1	Dermatology	34	14	20
2	Heart	13	6	5
3	Primary	17	7	9

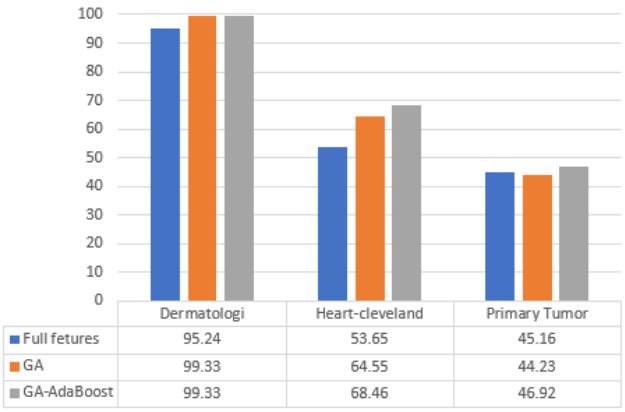

Figure 2: SVM Accuracy Performance as Evaluator Algorithm (%)

Table 3: Accuracy Classification Performance (%) using three classifiers: Naïve Bayes, k-NN, and RFT

Classifier	Dataset	full features	GA	GA-AdaBoost
Naïve Bayes	Dermatology	86.48	86.88	**88.19**
	Heart	25.77	47.12	**58.65**
	Primary Tumor	24.01	18.13	**25.38**
k-NN	Dermatology	**93.19**	87.81	88.19
	Heart	57.76	55.32	**57.88**
	Primary Tumor	31.76	33.52	**37.69**
Random Forest Tree	Dermatology	92.42	89.9	**92.67**
	Heart	56.99	56.99	**57.09**
	Primary Tumor	28.18	30.55	**31.21**

Table 4: Accuracy rank of each feature selection scheme using Wins-Loses method

Algorithm	Scheme	Wins	Loses	Wins-Loses
Naïve Bayes	**GA-AdaBoost**	**6**	**0**	**6**
	GA	1	4	-3
	Full features	1	4	-3
k-NN	**GA-AdaBoost**	**5**	**1**	**4**
	Full features	3	3	0
	GA	1	5	-4
Random Forest Tree	**GA-AdaBoost**	**6**	**0**	**6**
	GA	2	4	-2
	Full features	1	4	-3

and increases significantly on primary tumor dataset with 37.69% accuracy. On RFT classifier, GA-AdaBoost feature selection scheme insignificantly increases with 92.87% accuracy and significantly increases on heart and primary tumor dataset, as much as 57.09% and 31.21% of accuracy respectively.

The result of feature selection scheme performance comparison can be presented using rank and wins-loses terminology (Witten et al, 2011) . The terminology ‘wins’ shows the number of feature selection scheme performance that is statistically significant compared to other feature selection when certain classifier is used. Otherwise, ‘loses’ shows the opposite. The rank of each feature selection schemes is obtained based on the subtraction of ‘wins’ and ‘loses’. Table 4 shows accuracy rank of each feature selection scheme using wins-loses method.

Using Naïve Bayes, GA-AdaBoost scheme ranks the highest with six wins and without any loses. Using K-NN, GA-AdaBoost scheme has the highest rank with four wins and one loses. Using Random Forest Tree, GA-AdaBoost scheme has the highest rank with six wins and zero loses. The use of three classifiers shows that GA-AdaBoost scheme is superior to GA schemes that work individually without feature selection.

5. CONCLUSION

This research provides one example of application of feature selection method using wrapper based on ensemble learning with SVM as evaluator to overcome high dimensional data. Ensemble learning technique may maximize the performance of wrapper method. The result showed that combination of Genetic Algorithm selection feature and AdaBoost ensemble learning technique can provide statistically significant accuracy improvement and is superior on the use of Naïve Bayes, k-NN, and RFT classifier for three datasets. This proves that feature selection scheme proposed is superior than other scheme combinations. Besides, SVM as evaluator can perform smoothly, proven by accuracy improvement achieved by GA-AdaBoost and SVM scheme, superior to GA-SVM or solely SVM without feature selection. Even though feature selection using ensemble algorithm needs longer computation time, GA-AdaBoost scheme obtains better performance and outperforms individually-working feature selection. It is proven by higher wins acquired on three classifiers.

In future, feature subset selection may be applied in real world using hybrid or other methods by employing ensemble learning algorithm.

ACKNOWLEDGEMENT

This research is supported by Department of Informatics, Faculty of Communication and Informatics, *Universitas Muhammadiyah Surakarta*. Authors are grateful of the parties who have contributed to this research.

REFERENCES

[1] L. Yu and H. Liu, "**Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution**," *Int. Conf. Mach. Learn.*, pp. 856–863, 2003.

[2] V. Tan, P.N., Steinbach, M. & Kumar, ***Introduction to Data Mining*.** Pearson Education, Inc, 2006.

[3] G. Chandrashekar and F. Sahin, "A **survey on feature selection methods**," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014. https://doi.org/10.1016/j.compeleceng.2013.11.024

[4] J. Arunadevi and M. Josephine Nithya, "**Comparison of Feature Selection Strategies for Classification using Rapid Miner**," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 11, no. 9, pp. 5921–5925, 2016.

[5] R. Kohavi and H. John, "**Artificial Intelligence Wrappers for feature subset selection**," vol. 97, no. 97, pp. 273–324, 2011. https://doi.org/10.1016/S0004-3702(97)00043-X

[6] I. S. Thaseen and C. A. Kumar, "**Intrusion Detection Model Using fusion of PCA and optimized SVM**," *Contemp. Comput. Informatics (IC3I), 2014 Int. Conf.*, pp. 879–884, 2014. https://doi.org/10.1109/IC3I.2014.7019692

[7] S. Koul and R. Chhikara, "**A Hybrid Genetic Algorithm to Improve Feature Selection**," *Int. J. Eng. Res. Technol.*, vol. 4, no. 05, pp. 414–419, 2015. https://doi.org/10.17577/IJERTV4IS050496

[8] G. R. Kumar, G. R. Kumar, and G. A. Ramachandra, "**An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets International Journal of Advanced Research in An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Data**," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. January, pp. 272–277, 2016.

[9] M. Fernando, K. Halim, and G. Sanjaya, "**Optimization Features Using GA-SVM Approach**," *Int. J. Sci. Res.*, vol. 4, no. 9, pp. 193–197, 2015.

[10] Maryam, N. A. Setiawan, and O. Wahyunggoro, "**A hybrid feature selection method using multiclass SVM for diagnosis of erythemato-squamous disease**," *AIP Conf. Proc.*, vol. 1867, no. August, 2017. https://doi.org/10.1063/1.4994451

[11] A. Ozcift and A. Gulten, "**A Robust Multi-Class Feature Selection Strategy Based on Rotation Forest Ensemble Algorithm for Diagnosis of Erythemato-Squamous Disease**s," *J. Med. Syst.*, vol. 36, no. 2, pp. 941–949, Apr. 2012. https://doi.org/10.1007/s10916-010-9558-0

[12] D. Guan, W. Yuan, Y. K. Lee, K. Najeebullah, and M. K. Rasel, "**A review of ensemble learning based feature selection**," *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, vol. 31, no. 3. pp. 190–198, 2014. https://doi.org/10.1080/02564602.2014.906859

[13] A. Deshpande and R. Sharma, "**Multilevel ensemble classifier using normalized feature based intrusion detection system**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 5, pp. 72–76, 2018. https://doi.org/10.30534/ijatcse/2018/02752018

[14] Y. Saeys, T. Abeel, and Y. de Peer, "**Robust Feature Selection Using Ensemble Feature Selection Techniques**," in *Machine Learning and Knowledge Discovery in Databases*, 2008, pp. 313–325. https://doi.org/10.1007/978-3-540-87481-2_21

[15] N. Hoque, M. Singh, and D. K. Bhattacharyya, "**EFS-MI : an ensemble feature selection method for classification An ensemble feature selection method**," *Complex Intell. Syst.*, vol. 4, no. 2, pp. 105–118, 2018. https://doi.org/10.1007/s40747-017-0060-x

[16] A. V. S. Kumar, "**Ensemble Online Sequential Extreme Learning Machine and Swarm Intelligent Based Feature Selection for Cleveland Heart Disease Prediction System**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 6, no. 5, pp. 84–91, 2017.

[17] T. Zhang, "**Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models**," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1921–1928.

[18] P. Refaeilzadeh, L. Tang, and H. Liu, "**Cross-Validation**," in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 532–538.

[19] X. Wu *et al.*, ***Top 10 algorithms in data mining***. 2008. https://doi.org/10.1201/9781420089653

[20] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "**Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks,**" *Expert Syst. Appl.*, vol. 41, no. 4, Part 2, pp. 1937–1946, 2014. https://doi.org/10.1016/j.eswa.2013.08.089

[21] I. H. Witten, E. Frank, and M. A. Hall, ***Data Mining; Practical Machine Learning Tools and Techniques,*** Third Edit. Morgan Kauffman, 2011.