Breast Cancer Diagnosis Improvement Using Feature Selection

E. S. Wahyuni, N. A. Setiawan, H. A. Nugroho
Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia.

Abstract

The objective of this research is to improve the breast cancer diagnosis performance by applying feature selection methods to several classification algorithms. This study uses Wimconsin Breast Cancer Dataset (WBCD). Feature selection methods based on Rough Set (RS) and F-score (FS) are used for several classification algorithms, which are SMO (Sequential Minimal Optimization), Multi-Layer Perceptron (MLP), Naive-Bayes, C4.5, Instance Base Learning (IBK) and PART. This study uses 10-fold cross validation as an evaluation method. The results show that feature selection methods can improve diagnosis performance with a smaller number of features.

Keywords: Classification algorithm, Feature selection, F-score, Rough set

1. Introduction

Breast cancer is the most common cancer causing death in women. The frequency of this disease is relatively high both in the developed and less developed countries. It is estimated that more than 508,000 women died worldwide in 2011 due to breast cancer [1]. Based on current world breast cancer Age-standardized, Indonesia has mortality rates of 18.6 per 100,000 population in 2008 [2].

In line with the information technology advancement, particularly in the field of artificial intelligence, machine learning techniques which is considered as artificial intelligence methods are introduced to help and improve the automated detection capability. Therefore the machine learning techniques can be used to
assist medical experts and medical data can be analysed in a short time period as well as more detailed [1].

Many researchers used statistical and artificial intelligence techniques to predict breast cancer. The aim of these techniques are to specify identification of patients into a benign or malignant [2].

High-dimensional of medical data is one of the problems in the application of machine learning techniques because it makes a negative effect on the process of analysis. Dealing with the high-dimensional medical data, reducing the feature becomes very important. The advantages of feature reduction are as follows: (1) to avoid over-fitting, (2) to reduce the complexity of data analysis and (3) to improve data analysis performance [3].

One of the solutions to reduce the high-dimensional data is by using feature selection. Feature selection is part of the pre-processing phase in the classification process. Feature selection greatly affects the classification accuracy. The process of feature selection is proposed in many studies to improve accuracy of breast cancer diagnosis [4-7].

Chen et al. [8] used Rough set (RS) as feature selection. RS feature selection identifies significant features and eliminates irrelevant features to produce a good learning model. Therefore it reduces the dimensionality of data. Akay et al. [1] used F-score (FS) as feature selection. FS is a simple feature selection technique that measures discrimination two sets of real numbers. Feature having a low FS value is considered to have a low discriminatory capability and vice versa. Feature with a high FS value will have a high discriminative ability anyway.

In aforementioned works, Akay et al. [3] and Chen et al. [10] only tried on one classification algorithm, that is Support Vector Machine (SVM). Both of them have good accuracy after implementing feature selection and SVM, but it is not representative for the general FS and RS feature selections performance on other classification algorithms.

Therefore, in this paper a technique based on FS and RS feature selections is proposed on seven classification algorithms, these classification algorithms are SMO (Sequential Minimal Optimization), Multi-Layer Perceptron (MLP), Naive-Bayes, C4.5, Instance Base Learning (IBK) and PART. The results of classification algorithm are compared for performance analysis.

2. Basic concept theory

2.1 Feature selection

The basic concept of feature selection methods used in this research is described the following sub-section.

2.1.1 F-score

FS is a simple technique that measures the discrimination of two sets of real numbers. With training vectors \(x_k, k = 1, \ldots, m\), if the number of positive and
negative instance $n^+$ and $n^-$ respectively, then the FS of the $i^{th}$ feature is defined in (1).

$$F_i = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i \right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i \right)^2}{\frac{1}{n^+ - 1} \sum_{k=1}^{n^+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n^- - 1} \sum_{k=1}^{n^-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$  \hspace{1cm} (1)

where $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ respectively are the average of the $i^{th}$ feature of the whole, positive, and negative data sets; $x_{k,i}^{(+)}$ is the $i^{th}$ feature of the $k^{th}$ positive instance, and $x_{k,i}^{(-)}$ is the $i^{th}$ feature of the $k^{th}$ negative instance. Discrimination between positive and negative sets indicated by the numerator, and the denominator indicates one within each of the two sets. A feature that has a large f-score is the feature that is more discriminative. Therefore, we use this score as the feature selection criterion [9].

2.1.2 Rough set

RS theory is a new intelligent mathematical tool proposed by Pawlak [12] to deal with uncertainty and incompleteness. It is based on the concept of an upper and a lower approximation of a set, the approximation space and models of sets. The main advantage of RS theory is that it does not need any preliminary or additional information about data: like probability in statistics or basic probability assignment in Dempster Shafer theory and membership grade in fuzzy set theory. One of the major applications of RS theory is the attribute reduction that the elimination of attributes. The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Using the dependency degree as a measure, attributes are removed and reduced set provides the same dependency degree as the original [8].

2.2 Classification methods

2.2.1 SMO

SMO is an algorithm that breaks the large quadratic programming (QP) optimization problem of SVM. SMO uses the small QP problem, therefore it avoids using a time-consuming numerical QP optimization as an inner loop [13].

2.2.2 MLP

MLP algorithm is an algorithm which adopts the workings of human neural networks. This algorithm is well known because the learning process is able to do as directed. It consists of input, output, and a hidden layer. Learning algorithms are done using back propagation. Determination of the optimal weights will lead to the proper classification results [14].
2.2.1 Naive Bayes

Naive-Bayes classifier provides a simple approach, with a clear semantics, to representing, using, and learning probabilistic knowledge. This method is designed to be used in supervised induction tasks, the accuracy of predicting the class of the test sample and in which case training includes class information is the goal of performance [15].

2.2.2 C4.5

C4.5 is an algorithm for constructing decision trees are among the most well known and widely used of all machine learning methods. It is use the information entropy concept, making decisions is done by splitting each data attributes into smaller subsets in order to examine the entropy differences, and choose the attributes with the highest normalized information gain. The splitting stops when finding subset instances belong to the same class, and thus the leaf node gets created. If no leaf node is detected, J48 creates a higher up node decision based on the expected class value [16].

2.2.3 IBK

IBK based on their similarity. It is one of the most popular algorithms for pattern recognition. It is kind of lazy learning where the function is only approximated locally and all computations are deferred until classification process. An object is classified by a majority of its neighbours. K is a positive integer. A set of objects for which the correct classification is known are selected as the neighbours [17].

2.2.4 PART

PART is simplicity method for rule induction. It adopts the separate-and-conquer strategy in that it builds a rule, removes the instances it covers, and continues creating rules recursively for the remaining instances until none are left [18].

2.3 Performance evaluation criteria

Confusion matrix is a visualisation tool which is commonly used to present the accuracy of the classifiers. It is used to show the relationships between outcomes and predicted classes. Table 1 shows this confusion matrix.

| Actual | Predicted positive | Predicted negative |
|--------|--------------------|--------------------|
| positive | TP                 | FN                 |
| negative | FT                 | TN                 |

The accuracy is the proportion of the total number of predictions that were correct. It is determined using (2).
Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (2)

3 Methodology

3.2 Data set

The research material used in this study is the Wisconsin Breast Cancer database (WBCD) taken from the UCI machine learning repository, sourced from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [19].

This dataset is commonly used by researchers who are using Machine learning as a classification method of breast cancer, the dataset contains 699 samples taken from needle aspirates from human breast cancer tissue, which are 16 instances are having a missing value. Because of missing values found in a very small number compared to the overall number of data, the 16 instances are removed, so that the number of instances is used as 683.

The data sets show in Table 2 consists of nine features, each of feature is represented as an integer between 1-10.

| Label | Attribute                  | Domain |
|-------|---------------------------|--------|
| C1    | Clump Thickness           | 1-10   |
| C2    | Uniformity of Cell Size   | 1-10   |
| C3    | Uniformity of Cell Shape  | 1-10   |
| C4    | Marginal Adhesion         | 1-10   |
| C5    | Single Epithelial Cell Size | 1-10   |
| C6    | Bare Nuclei               | 1-10   |
| C7    | Bland Chromatin           | 1-10   |
| C8    | Normal Nucleoli           | 1-10   |
| C9    | Mitoses                   | 1-10   |

3.3 Research flow

3.3.1 Optimizing feature subset two approaches

RS feature selection, attribute reduction using Rough set theory in application ROSETTA [20], and attribute reduction is determined by choosing genetic algorithm (GA). A number subset of attributes obtained after reduction results, then selected the optimal subset attributes based on a subset of attributes that contain both attributes with strongest relevancy and weakest relevancy with decision attribute [21]. Selected optimum subset is shown in Table 3.
### Table 3: The selected attribute set based on Rough set

| No | Set attributes          |
|----|-------------------------|
| 1  | C2, C5, C6, C7, C9      |
| 2  | C2, C5, C6, C8, C9      |
| 3  | C5, C4, C6, C7, C9      |
| 4  | C1, C2, C5, C6, C9      |
| 5  | C1, C3, C4, C6, C9      |
| 6  | C1, C2, C4, C6, C9      |
| 7  | C5, C6, C7, C8, C9      |

FS feature selection, feature reduction is done by calculating the value of the FS for each attribute then the attribute value will be sorted from high to low then nine models with different number of attributes are composed, the nine subset attributes are shown in Table 4.

### Table 4: The nine feature subsets based on F-score

| No | Atribut       | Set Attributes |
|----|---------------|----------------|
| 1  |               | C6             |
| 2  |               | C6, C3         |
| 3  |               | C6, C3, C2     |
| 4  |               | C6, C3, C2, C7 |
| 5  |               | C6, C3, C2, C7, C1 |
| 6  |               | C6, C3, C2, C7, C1, C8 |
| 7  |               | C6, C3, C2, C7, C1, C8, C4 |
| 8  |               | C6, C3, C2, C7, C1, C8, C4, C5 |
| 9  |               | C6, C3, C2, C7, C1, C8, C4, C5, C9 |

### 3.3.2 Parameter setting

Several classification algorithms require the setting of specific parameters. SMO algorithm applied in this research using RBF (Radial Basis Function) kernel as the kernel function, there are two parameters, C and Gamma, that should be determined. This study applies grid search techniques to find the optimum parameters of C and gamma, with the grid space log₂ C \{5, 6, 7, ..., 20\} and log₂ γ \{-10, -11, ..., -5\}.

Multi-layer perceptron is applied in this research used a three-layer, this MLP consists of an input layer (28 neurons), one hidden layer (15 neurons) and one output layer (two neurons). Weight adjustment is done at 500 cycles. C.45 and PART algorithms applied in this research using a standard confidence factor (25%).
IBK algorithm applied in this research using 1 for the number of neighbors, and linear search as a nearest neighbor search algorithm.

3.3.3 Classification

The Weka data mining tool was used to evaluate the performance of SMO, MLP, Naive Bayes, C4.5, IBK and PART in each subset feature generated by two proposed features selection. 10-fold cross validation method was selected for performance evaluation of each classification algorithm. To make sure that the results were not biased, this research underwent 100 independent runs of the experiment and the average classification accuracies were computed, respectively. A scheme of this research can be seen in figure 1.

![Research scheme flow](image)

4 Result and discussion

4.2 Experiment (1) using Rough set feature selection.

Table 5 shows the comparison accuracy for each subset of attributes reduction results from RS feature selection with several classification algorithms (SMO, MLP, Naive Bayes, C4.5, IBK, and PART). The bold values are the highest accuracy.

The highest accuracy was obtained on a subset #5 (C1, C3, C4, C6, C9) by SMO and C4.5, with an accuracy of respectively 97.26% and 96.38%. MLP obtains the highest accuracy on the subset #4 (C1, C2, C5, C6, C9) 96.61%. PART obtains the highest accuracy on the subset #2 (C2, C5, C6, C8, C9) 95.87%. Meanwhile Naive Bayes and IBK obtain the highest accuracy on a subset #6 (C1, C2, C4, C6, C9), with an accuracy of 96.31% and 96.35%, respectively.
Table 5: Classification accuracies RS + several classification algorithms

| Subset | Classification accuracy (%) | SMO | MLP | Naive Bayes | C.45 | IBK | PART |
|--------|-------------------------------|-----|-----|-------------|------|-----|------|
| #1     |                              | 96.16 | 96.26 | 95.64 | 94.63 | 94.65 | 94.75 |
| #2     |                              | 96.19 | 96.42 | 95.88 | 95.59 | 95.90 | **95.87** |
| #3     |                              | 96.73 | 96.52 | 95.80 | 95.53 | 94.80 | 95.46 |
| #4     |                              | 96.57 | 96.61 | 95.74 | 95.57 | 95.98 | 95.76 |
| #5     | **97.26**                    | 96.53 | 96.30 | **96.38** | 95.75 | 95.57 |
| #6     |                              | 96.37 | 96.32 | **96.31** | 95.51 | **96.35** | 95.72 |
| #7     |                              | 96.01 | 95.80 | 95.24 | 94.75 | 95.05 | 95.11 |

4.3 Experiment (2) using F-score feature selection.

Table 6 shows the comparison accuracy of each subset of attributes reduction results from FS feature selection with several classification algorithms (SMO, MLP, Naive Bayes, C.45, IBK, and PART). The bold values are the highest accuracy.

The highest accuracy was obtained in a subset #7 (C6, C3, C2, C7, C1, C8, C4) by SMO, Naive Bayes, C4.5 and PART, with an accuracy of respectively 97.06%, 96.79%, 95.35%, and 95.76%. MLP obtains the highest accuracy on the subset #4 (C6, C3, C2, C7) 96.41%. Meanwhile IBK obtains the highest accuracy on the subset #9 (C6, C3, C2, C7, C1, C8, C4, C5, C9) 96.08%.

Table 6: Classification accuracies FS + several classification algorithms

| Subset | Classification accuracy (%) | SMO | MLP | Naive Bayes | C.45 | IBK | PART |
|--------|-------------------------------|-----|-----|-------------|------|-----|------|
| #1     |                              | 91.06 | 91.08 | 90.93 | 91.01 | 90.31 | 90.69 |
| #2     |                              | 95.86 | 95.95 | 95.24 | 94.90 | 94.69 | 95.28 |
| #3     |                              | 96.24 | 96.45 | 96.17 | 95.28 | 94.55 | 95.73 |
| #4     |                              | 96.30 | **96.41** | 95.41 | 95.19 | 94.59 | 95.48 |
| #5     |                              | 96.82 | 96.19 | 96.40 | 95.32 | 95.49 | 95.28 |
| #6     |                              | 96.99 | 96.13 | 96.52 | 95.18 | 95.82 | 95.37 |
| #7     | **97.06**                    | 95.82 | **96.79** | **96.35** | 95.72 | **95.76** |
| #8     |                              | 96.99 | 95.80 | 96.61 | 95.33 | 96.02 | 95.62 |
| #9     |                              | 97.01 | 95.88 | 96.13 | 95.28 | **96.08** | 95.63 |
4.4 Experiment (3) comparison result before and after implementing feature selection.

Table 7 shows the comparison of the classification performance before and after implementing feature selection. Overall, the performance of classification with feature selection has a higher accuracy than without feature selection. This proved feature reduction algorithm can improve the performance in classification. The bold values are the highest accuracy.

| Method       | Total attribute | Accuracy (%) |
|--------------|-----------------|--------------|
| SMO          | 9               | 97.01        |
| RS + SMO     | 5               | 97.26        |
| FS + SMO     | 7               | 97.06        |
| MLP          | 9               | 95.88        |
| RS + MLP     | 4               | 96.61        |
| FS + MLP     | 4               | 96.41        |
| Naive Bayes  | 9               | 96.13        |
| RS + Naive Bayes | 6             | 96.31        |
| FS + Naive Bayes | 7            | **96.79**    |
| C4.5         | 9               | 95.28        |
| RS + C4.5    | 5               | **96.38**    |
| FS + C4.5    | 7               | 95.35        |
| IBK          | 9               | 96.08        |
| RS + IBK     | 6               | **96.35**    |
| FS + IBK     | 9               | 96.08        |
| PART         | 9               | 95.63        |
| RS + PART    | 2               | **95.87**    |
| FS + PART    | 7               | 95.76        |

From table 7 it can be seen that the proposed method based SMO, the RS + SMO gives the best accuracy. For method based MLP, the RS + MLP gives the best accuracy. For method based Naive Bayes, the FS + Naive Bayes gives the best accuracy. For method based C4.5, the RS + C4.5 gives the best accuracy. For method based IBK, the RS + IBK gives the best accuracy. For method based PART, the RS + PART gives the best accuracy.

5 Conclusion

This research applied two feature selection methods which are RS and FS to several classification algorithms (SMO, MLP, Naive Bayes, C4.5, IBK and PART). The experimental results show that the implementation of feature selection process can generally improve all of the accuracy of the classification.
algorithm. Except for the method based on IBK, it can be seen from table 7 that the algorithm still gives the best accuracy without FS feature selection process.

References

[1] http://www.who.int/cancer/detection/breastcancer/en/index1.html.
[2] http://chartsbin.com/view/2sz.
[3] Akay, M. F., Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications, vol. 36, pp. 3240-3247, 2009.
[4] Soria, D., Garibaldi, J. M. E., Biganzoli, and Ellis, I. O., A Comparison of Three Different Methods for Classification of Breast Cancer Data, in Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on, 2008, pp. 619-624.
[5] Donghai, G., Weiwei, Y., Zilong, J., and Sungyoung, L., Undiagnosed samples aided rough set feature selection for medical data, in Parallel Distributed and Grid Computing (PDGC), 2012 2nd IEEE International Conference on, 2012, pp. 639-644.
[6] Moustakidis, S. P. and Theocharis, J. B., SVM-FuzCoC: A novel SVM-based feature selection method using a fuzzy complementary criterion, Pattern Recognition, vol. 43, pp. 3712-3729, 2010.
[7] Aijuan, D. and Baoying,W. Feature selection and analysis on mammogram classification, in Communications, Computers and Signal Processing, 2009. PacRim 2009. IEEE Pacific Rim Conference on, 2009, pp. 731-735.
[8] Hasan, H. and Tahir, N. M., Feature selection of breast cancer based on Principal Component Analysis, in Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium on, 2010, pp. 1-4.
[9] Osareh, A. and Shadgar, B., Machine learning techniques to diagnose breast cancer, in Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on, 2010, pp. 114-120.
[10] Chen, H.-L., Yang, B., Liu, J., and Liu, D.-Y., A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, Expert Systems with Applications, vol. 38, pp. 9014-9022, 2011.
[11] Chen, Y. W. and Lin, C. J., Combining SVMs with Various Feature Selection Strategies.
[12] Pawlak, Z., Rough sets, International Journal of Parallel Programming, 1982, 11 (5), 341-356.
[13] Platt, J. C., Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.
[14] Muliantara, A. and Widiartha, I. M., Penerapan multi layer preceptron dalam anotasi image secara otomatis.
[15] John, G. H. and Langley, P., Estimating Continuous Distributions in Bayesian Classifiers, Proceedings of the eleventh conference on Uncertainty in artificial Intelligence, 1995.
[16] Salzberg, S. L., Book Review: C5.5: Programs for Machine Learning, *Machine Learning*, 1994.

[17] Salama, G. I., Abdelhalim, M. B., and Zeid, M. A., Experimental comparison of classifiers for breast cancer diagnosis, in *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*, 2012, pp. 180-185.

[18] Frank, E. and Witten, I. H., Accurate Rule Sets Without Global Optimization, In: *Fifteenth International Conference on Machine Learning*, pp. 144-151, 1998.

[19] http://archive.ics.uci.edu/ml/datasets.html.

[20] http://www.lcb.uu.se/tools/rosetta/.

[21] John, G. H., Kohavi, R., and Pfleger, K., Irrelevant Features and the Subset Selection Problem, 1994.