Feature Selection Using Rough Set Theory Algorithm for Breast Cancer Diagnosis

D N K Hardani¹*, and H A Nugroho²

¹Department of Electrical Engineering, Universitas Muhammadiyah Purwokerto
²Department of Electrical and Information Engineering, Universitas Gadjah Mada

*e-mail address: diannova.kh@ump.ac.id

Abstract. Feature selection is one of the pre-processing stages of classification carried out by selecting relevant features that affect the results of classification. The advantage of feature selection is that it increases the value of accuracy. Data mining in the medical world has excellent potential for knowing hidden patterns in medical data sets. However, medical data sets often have large dimensions and have irrelevant features that can decrease the performance of the algorithm. This study aims to analyse the performance of the rough set approach as an algorithm used for feature selection in breast cancer diagnosis cases. This study conducted a feature selection process on the Wisconsin Breast Cancer (Diagnostic) Data Set provided by the UCI machine learning repository. There are several steps taken in research to realize these goals, such as data pre-processing, feature selection, data randomization, classification and performance evaluation. The result shows that feature selection using the rough set of methods has proven to be effective in reducing a large number of features in the data set.

1. Introduction

Cancer is a condition which cells have lost control and standard mechanisms so that they experience abnormal growth, fast and uncontrolled. Cancer can begin to grow in the mammary glands, milk ducts, fat tissue, or connective tissue in the breast. The key to surviving breast cancer survivors is to detect breast cancer as early as possible before cancer has a chance to spread. The breast cancer diagnosis method used by a doctor depends on the doctor's knowledge, intuition, and experience. The results of the diagnosis can affect the recommendation to the patients.

As technology advances, especially in the area of artificial intelligence, machine learning techniques are introduced to help improve automatic detection capabilities. With the help of this system, the possibility of misdiagnosis made by experts can be avoided, and medical data can be examined in a short time and more detail. The diagnosis of using computer-aided utilizes many data mining methods. Data mining with intelligent algorithms can be used to overcome diagnostic problems with medical data sets that involve multiple data inputs [1]. Data mining in the medical world has excellent potential for knowing hidden patterns in medical data sets. However, medical data sets often have large dimensions and have irrelevant features that can decrease algorithm performance. Feature selection is a pre-processing stage of classification carried out by selecting relevant features to improve the effectiveness and efficiency of the performance of the classification algorithm [2]. The advantage of feature selection is that it increases the value of accuracy [3].

Research on feature selection has been carried out in various cases. Feature Selection research for breast cancer diagnosis uses a wrapper approach based on genetic algorithm (GA) and case-based reasoning (CBR). GA is used to find the problem space that is aimed to find all possible subsets of features, and CBR is used to estimate the evaluation of the results in each subset. The experimental
results show that the proposed model can be compared with other models in the Wisconsin breast cancer data set (WDBC) [4]. Research using illustrative case studies in the Indian cosmetics industry illustrated the advantages of a rough set approach compared to conventional techniques for extracting decision rules from data sets, which can be useful in a variety of marketing applications. The rule generated through the methodology can perform as an expert, which may be referred to in future strategic decision-making. The approach gave similar results with the results obtained by statistical methods but without making any assumptions [5]. Ruhe [6] applied a rough set approach to analyse software engineering data that is resulted from a goal-oriented measurement. One method of feature selection that has been widely used and has a high degree of accuracy is the rough set theory approach. The attribute reduction for a decision system contains many decision attributes [7]. Other research work proposed a reasonable definition of parameterization reduction of soft sets and compared it with the attribute reduction in Rough Set Theory [8]. Zeng [9] conducted the development of a noise source classification system based on new methods for feature selection. The method used was a new rough set-based feature selection. It successfully reduced the number of features and improved classification accuracy.

This study aims to analyse the performance of the rough set approach as an algorithm used for feature selection in breast cancer diagnosis cases. The rough set approach was chosen because it can identify significant features and eliminate features that were not relevant to produce a good learning model and to reduce the dimensions of the data without the lack of information contained in the data set.

2. Methods
This study performed a feature selection process in the Wisconsin Breast Cancer (Diagnostic) Data Set provided by the UCI machine learning repository [10]. The data set contained of 32 attributes, 569 instances and two classes. The number of instances consisted of 357 benign and 212 malignant. These attributes were computed from a digitized image of a fine needle aspirate (FNA) of breast mass. The attributes information is shown in Table 1. Ten real-valued features are computed for each cell nucleus.

| No. | Attribute/Feature                                      |
|-----|--------------------------------------------------------|
| #1  | ID number                                              |
| #2  | Diagnosis                                              |
|     | Radius (mean of distances from center to points on the  |
|     | perimeter)                                             |
|     | Texture (standard deviation of gray-scale values)      |
| #3  | Perimeter                                              |
|     | Area                                                   |
|     | Smoothness (local variation in radius lengths)         |
|     | Compactness (perimeter^2 / area - 1.0)                 |
| #32 | Concavity (severity of concave portions of the contour) |
|     | Concave points (number of concave portions of the contour) |
|     | Symmetry                                               |
|     | Fractal dimension (“coastline approximation” - 1)      |

There are several steps taken in this research work to realize these goals. Figure 1 shows an overview of the research flow.

2.1. Data Preprocessing
Pre-processing data is the first step in diagnosing breast cancer. Data pre-processing conducted in this research is numerical data discretization. Some classification algorithms can provide better performance results when the data has been discretized. This research used the Boolean reasoning algorithm as a method of discretization process [11][12] and Rosetta toolkit [13].
2.2. Feature Selection

The feature selection process carried out to reduce the number of features in the WBCD data set and select features that are relevant to the results of breast cancer diagnosis decisions. In this study, two types of feature selection methods were used, namely Rough Set Theory (RST)-based feature selection and Correlation Based Feature (CFS)-based feature selection.

![Research flow diagram]

**Figure 1.** Research flow

RST provided a minimum set of features that can distinguish objects as well as the entire set of features used. However, the minimum feature set may not necessarily provide optimal performance when the classification process was carried out. In the rough set-based feature selection, the method used to reduce data is based on Johnson's algorithm so that a minimum reduction was obtained [14]. The feature selection process using the RST algorithm is shown in Figure 2.

CFS assumed that a subset of features containing of features that highly correlated with class and that each uncorrelated feature can improve classification performance. However, this subset of features was not necessarily a minimal set of features, such as reduction generated by RST. CFS was combined with the best first search strategy to select the correlation-based features. The feature selection process using the CFS algorithm is delineated in Figure 3. After the feature selection process was complete, a data set of some of the attributes selected in the feature selection process is obtained.

2.3. Data Randomization

The data randomization process in a data set can affect the performance of classification because the classification model formed during the training was different. Therefore, when the model tested on testing data, it would give different performance values. This data randomization process carried out using 10-fold cross-validation with ten iterations, any 100 data scrambling performance obtained for each classifier.
2.4. Classification
The computer-based breast cancer diagnosis was done by classifying the WBCD dataset using a classification algorithm. Classification performance was presented by performing 10-fold cross-validation on the randomized data set. This method was a common method used in data mining research. 10-fold cross-validation divided the data set into ten parts (fold). In this study, the classifiers used to classify data sets were Naive Bayes, Deep Learning, Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO), and J48. Every classifier used in this study followed the default settings and parameter provided by the Weka software [15].

2.5. Performance Evaluation
A comparison of classifier performance was presented by using three performance matrices, namely accuracy, sensitivity and specificity. Analysis of accuracy, sensitivity, and specificity can provide information about learning scheme trends that produce the best performance.

3. Result and Discussion
Based on the steps that have been taken for feature selection, the results obtained for each stage are described in this following section.

3.1. Data Discretization
Data discretization process was used to change the attributes of a numeric type data set to discrete. Some numeric data type attributes owned by the WBCD data set are converted to discrete type attributes using the Boolean reasoning algorithm discretization method. The results of the discretization are shown in Table 2.

Table 2. Discretization Result Value

| id        | texture_mean | texture_se | area_se | symmetry_se | perimeter_worst | smoothness_worst | Concavepoints_worst |
|-----------|--------------|------------|---------|-------------|-----------------|------------------|---------------------|
| [*, 87905) | [* 16.68]    | [* 0.8941] | [* 33.01] | [* 0.01649] | [* 105.95]     | [* 0.1256]       | [* 0.1358]         |
| [87905, 892202] | [16.68, 19.66] | [0.8941, 1.3605] | [33.01, *) | [0.01649, *) | [105.95, *) | [0.1256, *) | [0.1358, *) |
| [892202, *) | [19.66, *) | [1.3605, *) | | | | | |

Figure 2. The feature selection process using RST method
Figure 3. The feature selection process using CFS method
Table 2 above shows the discretization process divided attributes at specific intervals. From 32 existing attributes, only 8 attributes discretized. The id attribute, texture_mean, and texture_se were discretized into 3 (three) intervals while the attributes area_se, symmetry_se, perimeter_worst, smoothness_worst, and concavepoints_worst divided into 2 (two) intervals. The ‘*’ mark in the result of discretization, which states that the value is ‘less than’ if the position on the left. While the position on the right represents the value ‘more than’. If there is no ‘*’ sign that the discrete value is ‘between’. For example, the texture_mean feature has a discrete value \([*, 16.68)\) meaning that the texture_mean value < 33.01, \([16.68, 19.66)\) means 16.68 ≤ texture_mean < 19.66, and the value \([19.66, \ast)\) means texture_mean ≥ 19.66.

3.2. Feature Selection
Determination of training data in the Rough Set Theory (RST) based feature selection process using Johnson’s Algorithm. This algorithm was very fast because it used a greedy search to find one reduction (or one reduction per object in the case of object-related reduction). Based on the attribute reduction process, the results show that 3 (three) attributes selected with 480 rules. Table 3 shows the results of the feature selection.

| No | Attribute      |
|----|----------------|
| 1  | Radius_mean    |
| 2  | Perimeter_mean |
| 3  | Area_mean      |

In the feature selection process using the CFS algorithm combined with the best first search strategy was obtained 6 (six) features as shown in Table 4. The selection method used 10-fold cross-validation. This feature selection was processed based on correlation values. The features will be selected if have a high correlation of decision attributes and a low correlation between features.

| Number of folds | %  | Number of Attributes | Attribute            |
|-----------------|----|----------------------|----------------------|
| 10              | 100| 7                    | compactness_mean     |
| 10              | 100| 21                   | fractal_dimension_se |
| 10              | 100| 26                   | smoothness_worst     |
| 10              | 100| 27                   | compactness_worst    |
| 10              | 100| 28                   | concavity_worst      |
| 10              | 100| 30                   | symmetry_worst       |

Table 3 shows that the RST-based feature selection method can choose fewer features than the CFS-based feature selection method in Table 4. If the computer-based feature selection method combined between RST and CFS algorithm will produce 9 (nine) features that can be used as input for the classification process to diagnose breast cancer.

3.3. Performance Evaluation
After conducting the randomization process by using 10-fold cross-validation with ten iterations, it produced a new data composition that was continued to classification process by using some classifiers such as Naïve Bayes, Deep Learning, MLP, SMO, and J48. The performance evaluation of each classifier is shown in Figures 4, 5 and 6 below.
Figure 4. Accuracy of Classification

Figure 5. Sensitivity of Classification

Figure 6. Specificity of Classification
Accuracy and sensitivity show the ability of the system to diagnose breast cancer. A classification system must be measured its performance so that the next step is to evaluate the performance of the classification method. Figures 4 and 5 show that the performance of the RST algorithm is more powerful than CFS algorithm. However, when the two algorithms were combined, the results show a significant improvement in classification performance. The ability of the system to predict conditions that are not true; in general, the RST algorithm has the most backward performance, as shown in Figure 6.

4. Conclusion
Feature selection using the rough set of methods has been proven to be effective in reducing a large number of features in the data set. It is beneficial for reducing computing time and improving the performance of the classification process. This system can improve the quality of life by predicting cancer in early development.

5. References
[1] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 3240–3247, 2009.
[2] A. Kustiyo, H. N. Firgiani, and E. P. Giri, “Seleksi Fitur Menggunakan Fast Correlation Based Filter pada Algoritma Voting Feature Intervals 5,” J. Ilm. Ilmu Komput., vol. 6, no. 2, 2008.
[3] K. Thangavel and A. Pethalakshmi, “Dimensionality reduction based on rough set theory: A review,” Appl. Soft Comput., vol. 9, no. 1, pp. 1–12, 2009.
[4] M. Darzi, A. Liae, M. Hosseini, and H. Asghari, “Feature selection for breast cancer diagnosis: A case-based wrapper approach,” World Acad. Sci. Eng. Technol., vol. 53, pp. 1142–1145, May 2011.
[5] S. Mahapatra, Sreekumar, and S. S. Mahapatra, “Attribute selection in marketing: A rough set approach,” IIMB Manag. Rev., vol. 22, no. 1, pp. 16–24, 2010.
[6] G. Ruhe, “Rough set-based data analysis in goal-oriented software measurement,” in Proceedings of the 3rd International Software Metrics Symposium, 1996, pp. 10–19.
[7] H. Li, X. Zhou, J. Zhao, and D. Liu, “Attribute Reduction in Decision-Theoretic Rough Set Model: A Further Investigation BT - Rough Sets and Knowledge Technology,” 2011, pp. 466–475.
[8] D. Chen, E. C. C. Tsang, D. S. Yeung, and X. Wang, “The parameterization reduction of soft sets and its applications,” Comput. Math. with Appl., vol. 49, no. 5, pp. 757–763, 2005.
[9] X. Zeng and Y. Zhan, “Development of a noise sources classification system based on new method for feature selection,” Appl. Acoust., vol. 66, no. 10, pp. 1196–1205, 2005.
[10] UCI, “Breast Cancer Wisconsin (Diagnostic) Data Set,” 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. [Accessed: 21-Aug-2019].
[11] H. S. Nguyen, “Approximate Boolean Reasoning Approach to Rough Sets and Data Mining BT - Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing,” 2005, pp. 12–22.
[12] A. Øhrn, “Discernibility and Rough Sets in Medicine: Tools and Applications,” Nor. Univ. Sci. Technol., 2000.
[13] M. Kierczak, “ROSETTA A Rough Set Toolkit for Analysis of Data,” The Linnaeus Centre for Bioinformatics, Uppsala University, SWEDEN, 2009. [Online]. Available: http://bioinf.icm.uu.se/rosetta/.
[14] D. S. Johnson, “Approximation algorithms for combinatorial problems,” J. Comput. Syst. Sci., vol. 9, no. 3, pp. 256–278, 1974.
[15] “WEKA The workbench for machine learning.” Department of Computer Science at the University of Waikato, New Zealand. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/.