Comparative Analysis of Classification Algorithms on Endometrial Cancer Data

A. Hency Juliet* and R. Padmajavalli
Research and Development Centre, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; hencyjuliet@gmail.com, padmahari2002@yahoo.com

Abstract

**Objective:** To expose the Performance of classification algorithms on endometrial cancer data. The best algorithms are listed based on the result of various test options and ranked based on their accuracies. **Methods and Analysis:** Classification is one of the data mining techniques used to find a model that describes the data classes or concepts. The class-label of strange instance is predicted with the help of classification. It compares the classification algorithms by measuring accuracies, speed and strength of algorithms using WEKA tool. Accuracies of classification algorithms are calculated by means of four different options. The error rate and time taken to build the model also measured. **Findings:** The accuracies of sixteen algorithms are measured by training set, test set, tenfold cross validation and percentage split testing options. The average accuracies are calculated, then compared and ranked with highest accuracy first. The best five algorithms are taken for final performance on endometrial cancer dataset. The accuracy of Random Forest algorithm is high, but it took 0.16 sec to build the model, whereas the IBK, Random Tree and KStar algorithms' performs well with 0sec to build the model. Bagging algorithm takes more time to build the model. In terms of time and accuracy IBK produces better results as compared to other algorithms. Random Forest algorithm is most excellent in provisos of correctly classified occurrence. **Novelty/Improvement:** With the 315 instances of endometrial cancer data, the time taken to build the model is zero for IBK, KStar and Random Tree algorithms. If the number of instance increases then time also will increase.

**Keywords:** Classification Algorithms, Endometrial Cancer, IBK, KStar, Random Tree

1. Introduction

Data mining is the method of extracting the data from the huge dataset¹. Classification is one of the techniques in data mining to allocate objects to one of several pre-defined groups. Data Classification is a two step method consisting of knowledge step used to make a classification model and a categorization step used to calculate the class labels for a given data². It serves as a descriptive modeling, to distinguish between objects of unlike classes. A Classification model can also serve in predictive modeling, to calculate the class label of unidentified records. This process is mainly fitting for describing data sets with dual or diminutive types. It is a methodical approach to construct a classification models from the input data set. It includes Function, Bayesian, Meta-learning, Lazy, Rule-Based, Decision Tree and Miscellaneous classifiers. Each method utilizes a learning algorithm to recognize a model that best fits the liaison between the attribute set and class label of the input data. An important point of the learning algorithm is to construct the representation with generalization facility i.e., the representation precisely forecast the class labels of formerly unidentified instances³.

Endometrial cancer is a cancer that takes place from the endometrium, that is, the inside layer of the uterus or womb. It is the effect of the irregular progress of cells that have the ability to occupy or spread to other parts of the body⁴. During the premature section of the sequence, before the ovaries release an egg, the ovaries form hormones known as estrogens. Estrogen causes the endometrium to condense so that it could cultivate an embryo if pregnancy occurs. A woman's hormone stability took a...
part in the growth of most endometrial cancers. Many of
the risk factors for endometrial cancer influence estrogen
levels\(^5\).

This section reviews the related work on cancer data
using data mining algorithms.

In\(^6\) carried out a comparative study on clustering can-
cer genetic material data, and anticipated new clustering
methods that get advantage of individuality of the gene
expression data, the therapeutic society used the “classic”
clustering methods.

In\(^7\) evaluate the performance using performance met-
rics furthermore they proved the classifiers efficiency for
the prediction of cancer and heart disease in diabetic
patients. The investigational outcome shows that the clas-
sification accuracy is superior to existing approaches.
The proposed approach gives high efficiency and reduces
complication. The algorithm performs well and classifies
the dataset well compared to conventional methods.

In\(^8\) introduced a new classification approach that
uses classification techniques and association rule min-
ing which are integrated by the associative classification
process. Interesting relationships and correlation among
cancer and a set of practice which is useful for decision
building in any area of science above all in spatial epide-
miology was projected by association rule technique.

In\(^9\) methodically describe the appearance of endome-
trial cancer associated genes and analysis the functions,
pathways, and networks. The data were taken from the
PubMed database.

A Priyanga\(^10\) proposed the cancer prediction system,
by comparing its predicted results with patient’s prior
medical information and it was analyzed using WEKA.
The Classification algorithm \(J48\) and ID3 compared and
proved ID3 gives maximum accuracy on training data.

In\(^11\) discuss the different types of data mining classi-
fication algorithms accuracies based on type-2 diabetes
disease dataset perspective to Bangladeshi populations.
The different classification algorithms are compared by
computing accuracies, speed and robustness of the algo-

\section{Materials and Methods}

Information retrieval is the method used to collect the
data from the large database. Using resample filter from
supervised learning, the instances are taken for different
test options. The data set contains 315 instances. 60% of
instances are taken as training set. From the remaining
40% instances, 20% of instances are taken for cross valida-
tion and 20% taken for supplied test set. Then the learning
algorithms of different classifiers are used on the cancer
data. The Classification algorithms such as Naïve Bayes,
ByesNet, SMO, IBK, KStar, Bagging, AdaBoostM1, VFI,
OneR, JRip, Random Forest, Random Tree, \(J48\), FTree
and Multiclass classifier are used to categorize the occur-
rence. The accuracies of sixteen algorithms are measured
by training set, test set, tenfold cross validation and per-
centage split testing options. Then the average accuracies
are compared and ranked with highest accuracy first. The
best five algorithms are taken for final performance on
endometrial cancer dataset. The process flow of compara-
tive analysis is illustrated in Figure 1.
2.1 Dataset Description

The endometrial cancer dataset is collected from the Breslow, N.E. and Day, N.E. Cancer Dataset\(^2\). The dataset contains 315 instances and 10 attributes with data on cases and controls from the study of endometrial cancer as related to treatment with estrogens for menopausal symptoms and other risk factors. The data descriptions are given in Table 1. By using the machine learning data mining tool WEKA, the performance of algorithm was uncovered based on the correctly classified instances. The accuracy was found using the confusion matrix.

| Number | Name | Description | Codes/Range       |
|--------|------|-------------|-------------------|
| 1      | SET  | Matched set indicator | 1-63              |
| 2      | CASE | Case-control indicator | 0 = Control, 1 = Case |
| 3      | AGE  | Age in years | 55-83             |
| 4      | GALL | Gallbladder disease | 0 = No, 1 = Yes   |
| 5      | HYP  | Hypertension | 0 = No, 1 = Yes   |
| 6      | OB   | Obesity      | 0 = No, 1 = Yes; 9 = Unknown |
| 7      | EST  | Estrogen usage | 0 = No, 1 = Yes  |
| 8      | DOSE | Dose of conjugated | 0 = 0, 1 = 0.3, 2 = 0.301-0.624, 3 = 0.625, 4 = 0.626-1.249, 5 = 1.25, 6 = 1.26-2.50, 9 = Unknown |
| 9      | DUR  | Duration of estrogen use (months) | 0-95, 96=96+, 99=Unknown |
| 10     | NON  | Non-estrogen drug | 0 = No, 1 = Yes   |

2.2 Overview of Endometrial Cancer

The most common type of adeno carcinoma is referred as endometriosis cancer. Endometriosis cancers are created by cells in glands that appear a lot like the usual uterine inside layer (endometrial). There are two types of endometrial cancer\(^3\). Type 1 endometrial cancers are deliberation to be origin by surfeit estrogen. They developed from a typical hyperplasia sometimes, an abnormal overgrowth of cells in the endometrium. Type 1 cancers are typically not very destructive and are slow to widen to new tissues. Type 2 endometrial cancers create a small number of endometrial cancers. It doesn't seem to be origin by a large amount estrogen\(^4\).

2.3 Accuracy Measurement

The assessment of the performance of a classification model is based on the number of test records appropriately and wrongly predicted by the model\(^5\). These calculations are put into a table called confusion matrix. It provides the information needed to determine the performance of classification model. Summarizing the information in a single number would build it more suitable to compare the performance of different models\(^6\). This can be done by the performance metric accuracy.

\[
\text{Accuracy} = \frac{\text{Number of exact predictions}}{\text{Total number of predictions}}
\]

Similarly the performance of a model can be expressed in terms of its error rate.

\[
\text{Error Rate} = \frac{\text{Number of incorrect predictions}}{\text{Total number of predictions}}
\]

3. Results and Discussion

In this paper the resample filter is used to maintain the class distribution in the sub sample, to bias the class allocation towards an even distribution. Then the accuracies of sixteen classification algorithms have been measured using various test options of WEKA tool. It is ample software, written in the Java language, has been created to illustrate the thoughts called the Waikato Environment for Knowledge Analysis.

The accuracies are measured in terms of correctly classified instances. The error rates can be measured in terms of incorrectly classified instances. The accuracies are compared and ranked with highest accuracies first. The best five algorithms are taken for final performance on endometrial cancer dataset which contains 315 instances with 10 attributes. Figure 2 shows the correctly and incorrectly classified instances with confusion matrix of Random forest classification Algorithm.

Table 2 shows the accuracies of sixteen algorithms in various test options. Table 3 shows the time taken to build model for the top five algorithms. Table 4 shows Error rates of top five algorithms and their averages.
### Table 2. Accuracies of different classification algorithms and their averages

| Classification algorithms | Training data set | Test Data Set | Ten fold cross validation | Percentage split | Average percent |
|--------------------------|-------------------|---------------|--------------------------|------------------|-----------------|
| NaiveBayes               | 77.60%            | 84.38%        | 76.04%                   | 78.46%           | 79.12%          |
| BayesNet                 | 78.13%            | 84.38%        | 78.65%                   | 83.08%           | 81.06%          |
| SMO                      | 80.21%            | 76.56%        | 78.13%                   | 83.08%           | 79.49%          |
| IBK                      | 99.48%            | 85.94%        | 75.00%                   | 83.08%           | 85.87%          |
| KStar                    | 99.48%            | 87.50%        | 75.52%                   | 78.46%           | 85.24%          |
| Bagging                  | 85.94%            | 82.81%        | 76.04%                   | 81.54%           | 81.58%          |
| Multiclass classifier    | 80.21%            | 82.81%        | 76.56%                   | 86.15%           | 81.43%          |
| AdaBoostM1               | 81.25%            | 81.25%        | 78.13%                   | 83.08%           | 80.93%          |
| VFI                      | 78.13%            | 84.38%        | 76.56%                   | 80.00%           | 79.77%          |
| JRip                     | 80.21%            | 76.56%        | 79.17%                   | 83.08%           | 79.75%          |
| OneR                     | 80.21%            | 76.56%        | 80.21%                   | 83.08%           | 80.01%          |
| Random Forest            | 99.48%            | 90.63%        | 76.04%                   | 83.08%           | 87.31%          |
| Random Tree              | 99.48%            | 90.63%        | 74.48%                   | 76.92%           | 85.38%          |
| J48                      | 82.81%            | 81.25%        | 71.35%                   | 83.08%           | 79.62%          |
| FT                       | 81.77%            | 82.81%        | 77.08%                   | 81.54%           | 80.80%          |
| Logistic                 | 80.21%            | 82.81%        | 76.56%                   | 86.15%           | 81.43%          |

### Table 3. Time taken to build model of top five classification algorithms and their averages

| Classification Algorithms | Training data set in Sec | Test Data Set in Sec | Ten fold cross valid in Sec | Percentage split in Sec | Average Time Taken to Build Model in Sec |
|--------------------------|--------------------------|----------------------|-----------------------------|-------------------------|------------------------------------------|
| IBK                      | 0.00                     | 0.00                 | 0.00                        | 0.02                    | 0.0050                                   |
| KStar                    | 0.00                     | 0.00                 | 0.00                        | 0.00                    | 0.0000                                   |
| Random Forest            | 0.09                     | 0.16                 | 0.09                        | 0.13                    | 0.1175                                   |
| Random Tree              | 0.00                     | 0.02                 | 0.00                        | 0.02                    | 0.0100                                   |
| Bagging                  | 0.22                     | 0.06                 | 0.09                        | 0.06                    | 0.1075                                   |

### Table 4. Error rates of top five classification algorithms and their averages

| Classification Algorithm | Training data set | Test Data Set | Ten fold cross valid | Percentage split | Average Error Rate |
|--------------------------|-------------------|---------------|----------------------|------------------|-------------------|
| IBK                      | 0.52%             | 14.06%        | 25.00%               | 16.92%           | 14.13%            |
| KStar                    | 0.52%             | 12.50%        | 24.48%               | 21.54%           | 14.76%            |
3.1 Performance Evaluation

The classification accuracy can be measured with the help of appropriately and wrongly classified instances. In Table 1 the accuracies are specified in terms of appropriately classified instance. The wrongly classified instance implies the error rate, which is the remaining percentage of correctly classified instance. Figure 3 shows the highest accuracies using bar graph. Here it is seen that, in training dataset four algorithms are parallel high in case of accuracy. Similarly in 10 fold cross validation two algorithm, in percentage split ten algorithms and in average one algorithm carry high accuracy.

Finally most five algorithms for Endometrial Cancer data are ranked. This work is also checked for speed and error rate. Random Forest, IBk, Random Tree and KStar algorithms have best accuracies for total training dataset. On the other hand Random Forest and Random Tree are best for test data case. OneR is best for 10 fold cross validation and Multiclass Classifier, Bagging algorithms are best for percentage split respectively. But Random Forest is best for all cases whereas IBk, Random Tree and KStar algorithms are next top ranker classification algorithms. Figure 3 shows the highest accuracies in different test options. Table 5 shows the average accuracies and average time taken to build model. Figure 4, 5 and 6 shows the graphical representation of average error rates, final performance and time taken to build model.

| Classification algorithms | Average Accuracy percentage | Average Time Taken to Build Model in Sec |
|---------------------------|-----------------------------|----------------------------------------|
| IBK                       | 85.87%                      | 0.0050                                 |
| KStar                     | 85.24%                      | 0.0000                                 |
| Random Forest             | 87.31%                      | 0.1175                                 |
| Random Tree               | 85.38%                      | 0.0100                                 |
| Bagging                   | 81.58%                      | 0.1075                                 |

Figure 3. Highest accuracies in different test options.

Figure 4. Average error rates of top five.
Comparative Analysis of Classification Algorithms on Endometrial Cancer Data

4. Conclusion

In this paper, comparative study has been performed on the algorithms such as Naïve Bayes, ByesNet, SMO, IBK, KStar, Bagging, AdaBoostM1, VFI, Decision Table, OneR, JRip, Random Forest, Random Tree, J48, FTree, Multiclass Classifier to classify the instances on endometrial cancer dataset and the comparison outcomes are presented in the outward appearance of graph and table. The comparative study is performed on the basis of accuracies of correctly classified instance. The accuracy of Random Forest algorithm is high, but it take 0.16 sec to build the model, whereas the IBK, Random Tree and KStar algorithms’ performs well with 0sec to build the model. Bagging algorithm takes more time to build the model. In terms of time and accuracy IBK produces better results as compared to other algorithms. In terms of correctly classified instance, Random Forest algorithm is best.

5. References

1. Jayaraj V, Mahalakshmi V. Augmenting efficiency of recruitment process using IRCF text mining algorithm. Indian Journal of Science and Technology. 2015 Jul; 8(16).
2. Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd ed. Simon Fraser University.
3. Karamizadeh F, Zolfagharifar SA. Using the clustering algorithms and rule-based of data mining to identify affecting factors in the profit and loss of third party insurance, insurance company auto. Indian Journal of Science and Technology. 2016 Feb; 9(7).
4. Saksouk FA. Endometrial Carcinoma Imaging. Eugene CL, editor
5. A detailed guide – Endometrial cancer. Available from: www.cancer.org/cancer/endometrial cancer/cancer-risk-factors
6. de Souto MCP. Clustering cancer gene expression data: A comparative study. BMC Bioinformatics. 2008; 4(497).
7. Kalaiselvi C, Nasira GM. Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian Journal of Science and Technology. 2015; 8(14).
8. Shridhar R. Association rule-spatial data mining approach for exploration of endometrial cancer data. International Journal of Advanced Research in Computer Science and Software Engineering. 2013.
9. Gao H, Zhang Z. BioMed Research International. 2015.
10. Priyanga A. Effectiveness of data mining - based cancer prediction system. International Journal of Computer Applications. 2013 Dec; 83(10). (0975 – 8887).
11. Ahmed K, Jesmin T. Comparative analysis of data mining classification. Internat J Sci Eng. 2014; 7. Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach.
12. The cancer genome atlas. A Pilot Project of the National Cancer Institute.
13. The cancer genome atlas research network. Integrated Genomic Characterization of Endometrial Carcinoma. Nature. 2013 May 2. DOI: 10.1038/nature12113.
14. A detailed guide – Endometrial cancer. Available from: http://www.cancer.org/cancer/endometrialcancer/what-is-endometrial-cancer
15. Gharehchopogh FS, Khaze SR, Maleki I. A new approach in bloggers classification with hybrid of K-nearest neighbor
and artificial neural network algorithms. Indian Journal of Science and Technology. 2015 Feb; 8(3):237–46.
16. Classification basic concepts. E-Book. Ch 4. p. 145–9.