A Comparative Study of Data Mining Methods to Diagnose Cervical Cancer

Noviyanti T M Sagala
Department of Information System, Krida Wacana Christian University – Indonesia

* noviyanti.sagala@ukrida.ac.id

Abstract. Cervical cancer becomes a major cause of cancer deaths in women around the world. The objective of this study is to provide a comprehensive analysis of different data mining methods to diagnose the malignant cancer samples. Different data mining algorithms (SVM, Naïve Bayes, and KNN) has been applied on four different medical tests (Biopsy, Cytology, Hinselmann, and Schiller) as four different target variables. The attributes influence the disease most is extracted since the disease has no symptoms in the early stage. The extraction involved over 32 attributes and two different algorithms such as Correlation-based Filter (CFS) and Random Forest. The results showed that the performance of Naïve Bayes classifier outperforms other classifiers after evaluation using 10-fold cross-validation method in R environment. In addition, the use of attribute selection has been proved not only can select the highly important attributes but also to increase the performance of all classifiers on cervical cancer dataset. In this study, the work reveals the classifiers can effectively achieve the best performance with the least number of highly important attributes.

1. Introduction

In recent years, cervical cancer is a disease that has a tendency to revive and becomes cancer-correlated death in the worldwide. The disease affects women who are less than 50 years of age [1], [2]. It is due to the abnormal growth of cells that have the power to damage nearby body parts such as the bladder, rectum, lungs, vagina, and liver.

Cervical cancer could be detected as it has developed to risky stage some of which are vaginal bleeding and cervical pain. In addition, the poor problem awareness of people to a periodic screening as well as the shortage of expertise and healthcare infrastructure bring about cervical cancer as a lead of mortality in developing countries [3]. Moreover, most of the expertise in diagnosing cervical cancer disease based on their learning and experience. The fundamental of effectual medical care should be performed at the early diagnosis. There are several risk factors of Cervical Cancer including HPV(Human Papilomavirus), Early Sexual Activity, Irregular sexual Contact, Long-term use of the contraceptive pill, having sexually transmitted infections (HIV/AIDS). Those factors may be collected from the survey, questionnaire, and medical record and be accurate to be used either in real life or in research work [4].

Data mining is considered as one of the major challenging and top major research field in medical due to the high importance of healthcare issues. In the medical area, data mining techniques are beneficial not only in finding patterns and relationships among some symptoms...
but also in predicting various diseases [5]. We came to know that by implementing different data mining methods continuously analysis and medical care can be suggested immediately which lead to even save the lives, especially the one who suffer cervical cancer. The first task is data preprocessing which plays essential roles as about eighty percent of the whole data mining task. Data preprocessing will assist to improve the model and decrease computational time in classifying the case.

The main direction of this work is to build a prediction cervical cancer model with a specific set of risk factors from four different clinical test results by applying Data Mining techniques. The study will discuss the comparison of the performance three different classification methods (KNN, Naïve Bayes, and SVM) with/without applying data reduction on cervical cancer data.

2. Literature Review

The previous works on the same dataset are described as following: Classification methods namely Naïve Bayes, J48 Decision Tree, Sequential Minimal Optimization, and Random Forest are presented and assessed on several parameters (exact match, ranking loss performance, accuracy, and hamming loss) in WEKA environment [6]. Risk factors of cervical cancer are reviewed with the combination of RFE and PCA with SVM method. The data reduction methods reduce the variables from 30 to 8 and drop the classification speed. In general, SVM-PCA achieved better performance compared to SVM-RFE and standard SVM [7].

The other works on different cervical cancer data showed the importance of data mining technique in real life applications. Classification results showed that 20 records with High risk, 20 records with moderate risk while 10 records with low risk [4]. Several data mining methods have been extensively performed to build an efficient prediction model of heart disease. Support Vector Machine model is the best prediction model among the other models built by Decision Tree and Artificial Neural Network [8]. Other work demonstrated that the prediction model for heart disease built by applying Naive Bayes technique is superior to Decision Tree or Neural Network [9]. The performance of KNN achieved higher accuracy than neural network ensemble and even the use of a voting technique could not improve the performance of KNN [10].

3. The Proposed Methodology

3.1. Dataset

Data was from UCI Machine Learning repository [11] which contains 858 records with 32 attributes and 4 target variables which indicate the type of medical test specifically; Hinselmann, Schiller, Cytology, and Biopsy. The information for attributes is described in Table 1 [12].

| No | Attribute Name                        | No | Attribute Name                                      |
|----|---------------------------------------|----|-----------------------------------------------------|
| 1  | Age                                   | 19 | STDs: pelvic inflammatory disease                   |
| 2  | Number of sexual partners             | 20 | STDs: genital herpes                                |
| 3  | First sexual intercourse(age)         | 21 | STDs: molluscum contagiosum                         |
| 4  | Number of pregnancies                 | 22 | STDs: AIDS                                          |
| 5  | Smokes                                | 23 | STDs: HIV                                           |
| 6  | Smokes(years)                         | 24 | STDs: Hepatitis B                                   |
| 7  | Smokes(packs/year)                   | 25 | STDs: HPV                                           |
| 8  | Hormonal Contraceptives              | 26 | STDs: Number of diagnosis                           |
| 9  | Hormonal Contraceptives(years)       | 27 | STDs: Time since first diagnosis                     |
| 10 | Intrauterine Devices(IUDs)           | 28 | STDs: Time since last diagnosis                      |
| 11 | IUDs(years)                           | 29 | Dx:Cancer                                           |
| 12 | Sexually Transmitted Diseases (STDs)  | 30 | Dx:CIN                                              |
| 13 | STDs (number)                        | 31 | Dx:HPV                                              |
| 14 | STDs:condylomatosis                  | 32 | Dx                                                   |
| 15 | STDs:cervical condylomatosis         | 33 | Hinselmann                                          |
| 16 | STDs:vaginal condylomatosis          | 34 | Schiller                                            |
3.2 Methodology

We have used the popular, open-source data mining tool RStudio Desktop 1.1.453 and R (version 3.4.3) for this analysis. The methodology follows the standard of data mining as shown in Figure 1.

![Figure 1. The Proposed Methodology](image)

- In data selection, it was observed that two attributes (STDs) were useless for data mining. Data reduction process points to two ways i.e. reduction the data dimensionality or reduction of the data distribution; both of the approaches were implemented on this work. Reducing the dimensionality of the data was performed by feature selection while reducing the distribution of the data was performed by data discretization. The purpose of data discretization is to diminish from a large domain of numeric values to a subset of categorical values and may help to improve the learning. The smaller data varieties normally contributed to more precise predictive model and bring out higher prediction rates into new cases [13]. The data discretization was only applied to several attributes through the use of library “discretization”. Attribute age was transformed into four discrete values according to [14] namely <21, 21-29, 30-65, and >65 years of age. Attribute number of sexual, first sex intercourse, number of pregnancies, smoke years, smoke packs, hormonal contraceptive years, and IUD years were transformed using MDLP into four intervals. The reduced attribute set was used to prepare training and testing dataset. Dimensionality reduction can be conducted with the help of Random Forest and Correlation-based Filter Selection. The different number of threshold was determined to select the best attribute set of Random Forest. There were 5, 10, 15, 20, and 25 top attributes.

- Classification

All attribute sets were given as input to the classifier to identify to which set of categories the input belongs to. Following are the classifiers used for classification in this work:

a. Support Vector Machine (SVM): SVM is well known as a classifier which has good accuracy and less prone to overfitting. The advantage of this algorithm is applied either to a linear or non-linear dataset.

b. Naive Bayes: Bayesian classifier is based on Bayes’ theorem. Theorem is follows [11]:

\[ P(T|Y) = \frac{P(Y|T) P(T)}{P(Y)} \]

In Bayesian, Y is considered as evidence and T be some hypothesis means, the data of Y belongs to particular class Z. To determine \( P(T|Y) \), the probability that the hypothesis T holds given evidence according to Bayes theorem the \( P(T|Y) \) is formulated as

\[ P(T|Y) = P(Y|T) P(T) / P(Y) \]

c. K-Nearest Neighbor (KNN): Compared to other machine learning algorithms, knn is the simplest. The method classifies targets on the closest k value of training samples in the attribute space. While in the process of classification, the same attributes are calculated for test data (which classification is undefined). The distance of this new vector is...
calculated using the Euclidean distance metric to all learning data vector. In this work, the good \( k \) value is selected with cross-validation.

- **Performance Evaluation**
  a. 10-fold Cross-validation
  This method splits the data set into 10 identical parts (“folds”), then uses apart as the testing and the remaining parts as the training. The operation is then rerun 10 times, using different folds as the testing set each time. The estimation accuracy of the predictive model was obtained from the average of all the rounds.
  b. Specificity is proportion of correctly diagnosed mild cancer pieces in all benign pieces.
  c. Sensitivity judges the ability of the model to precisely recognized diseased patients
- **Final Prediction Model** was selected based on the highest values of average accuracy, sensitivity, and specificity.

4. **Results and Discussion**
After removing incomplete records and two useless variables, 668 records with 30 attributes becomes the final data composition of data selection step. Then, eight variables were transformed into four intervals as shown in table 2.

| Name of attribute | Age | Number of sex partner | First sex intercourse | Number of pregnancies | Smokes years | Smokes packs years | Hormonal contraceptive years | IUD years |
|-------------------|-----|-----------------------|----------------------|-----------------------|-------------|-------------------|-----------------------------|-----------|
| Before discretization | 43  | 13                    | 22                   | 12                    | 31          | 63                | 41                          | 27        |
| After discretization | 4   | 4                     | 4                    | 4                     | 4           | 4                 | 4                           | 4         |

Random Forest and CFS were then applied on the discretized data to select highly important attributes. Since Random Forest works as multivariate approach (ranking), there are no absolute amount of selected attributes, therefore several number of threshold were determined to obtain the total amount of attribute sets. The result of Random Forest and CFS are shown in table 3 and table 4, respectively. Numbers in those tables are correspond to the numbers of risk factors in table 1.

| Target Variable | Rank of attribute (Decreasing order) |
|-----------------|--------------------------------------|
| H               | 4,9,1,2,11,3,26,31,13,6,29,7,8,11,5 |
| B               | 4,9,2,1,3,11,31,7,6,29,8,32,13,11,23 |
| C               | 9,4,2,1,3,11,8,7,32,13,23,11,6,5,29 |
| S               | 9,4,1,2,3,11,7,6,8,13,31,26,29,23,5,11 |

Table 3 describes that all target variable has similar rank of attributes especially on top six attributes. These include age, number of sexual partners, first sexual intercourse (age), number of pregnancies, hormonal contraceptive (years), and IUD (years). The higher the rank, the stronger the correlation of attribute to target variable, the higher the influence of the risk factor the patient suffered the disease. We formed the attribute sets based on the rank and number of threshold.

| Target Variables | Number of selected attribute | Corresponding number of attribute |
|------------------|------------------------------|----------------------------------|
| H                | 4                            | 2,3,4,11                         |
| B                | 4                            | 3,4,9,11                         |
| C                | 3                            | 7,9,11                           |
| S                | 1                            | 11                               |
From table 4, the total number of attribute in all target variable is less than five attributes even in Schiller test is only an attribute. In addition, attribute IUD (years) is considerably important according to the existence in the result. The total of seven attribute sets as input to classification process; full feature set, five feature sets of RF, and a feature set of CFS. Final classification method is set to the one that achieves the best performance.

- **Target Variable: Biopsy**
  The capability of NB-CFS scheme with four attributes is better than the other schemes. It is found that NB-RF with full feature set is the lowest accuracy. In spite of that, the model KNN-RF works as well as KNN-CFS/NB-CFS model. The attribute selection methods increases the performance of the classifiers.

| Table 5. Result for Biopsy |
|---------------------------|
|                           | Full Feature | RF | CFS |
|                           | NB | KNN | SVM | NB | KNN | SVM | NB | KNN | SVM |
| Total accuracy            | 87.97 | 90.46 | 95.20 | 92.48 | 96.24 | 94.24 | 96.24 | 96.24 | 93.24 |
| Specificity               | 85.72 | 89.20 | 90.20 | 89.20 | 92.50 | 90.10 | 92.50 | 92.50 | 89.45 |
| Sensitivity               | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

- **Target Variable: Cytology**
  NB-CFS achieves the same performance as KNN-RF. Although the performance of those schemes was same but NB-CFS worked faster than KNN-RF in terms of building predictive model. When applying the RF/CFS, the performance of KNN and NB improved. The result worked on contrary to SVM.

| Table 6. Result for Cytology |
|-----------------------------|
|                            | Full Feature | RF | CFS |
|                            | NB | KNN | SVM | NB | KNN | SVM | NB | KNN | SVM |
| Total accuracy             | 85.23 | 85.12 | 94.75 | 92.48 | 96.24 | 94.24 | 96.24 | 96.24 | 94.74 |
| Specificity                | 84.12 | 83.14 | 89.92 | 89.23 | 92.39 | 89.23 | 92.39 | 89.23 | 85.32 |
| Sensitivity                | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

- **Target Variable: Hinselmann**
  In General, the performance of Naïve Bayes is outperform other classifiers. Total accuracy of NB-RF is slightly higher than NB-CFS. It is inspected that KNN showed improvement on performance in smaller amount of attribute.

| Table 7. Result for Hinselmann |
|-------------------------------|
|                              | Full Feature | RF | CFS |
|                              | NB | KNN | SVM | NB | KNN | SVM | NB | KNN | SVM |
| Total accuracy               | 87.97 | 87.90 | 94.01 | 95.49 | 93.98 | 93.98 | 94.23 | 94.20 | 90.75 |
| Specificity                  | 85 | 85.20 | 87.20 | 91.1 | 89.03 | 89.03 | 91.41 | 88.12 | 84.68 |
| Sensitivity                  | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

- **Target Variable: Schiller**
  Compared with the other three target variables, Schiller’ test downfall to various test results. Specifically, utilizing only a feature to build the predictive model decreased the performance of all classifier. It is discovered that NB-RF is excellent than other classification schemes.

| Table 8. Results for Schiller |
|------------------------------|
|                             | Full Feature | RF | CFS |
|                             | NB | KNN | SVM | NB | KNN | SVM | NB | KNN | SVM |
| Total accuracy              | 85.45 | 84.12 | 90.23 | 92.48 | 89.47 | 89.47 | 90.98 | 89.47 | 86.47 |
| Specificity                 | 83.30 | 82.25 | 84.46 | 89.56 | 85.47 | 85.47 | 88.43 | 85.47 | 80.23 |
| Sensitivity                 | 96.31 | 96.23 | 97.23 | 98.46 | 97.21 | 97.21 | 97.46 | 97.21 | 98.70 |

It is marked that the performance of Naive Bayes with CFS achieves higher accuracy,
specificity, and sensitivity than either other classifiers with CFS or Naïve Bayes with/without feature selection methods. It is expected because Naïve Bayes may perform better on less number of relevant attribute. Oppositely, the performance of SVM increased when applying higher number of attribute. From the experimental results, we found KNN achieved the best performance at K=7. In terms of computational time, Naïve Bayes needs shorter time to build the model than other classifiers. The selected attributes from CFS and RF is also persistent to medical result evidences [1]. Furthermore, applying feature selection methods can help to speed up the achievement of classifier and slow down the computational time.

5. Conclusion
In this study, various data mining methods on cervical cancer data is performed. The results of data reduction helped to improve the performance of the methods. Generally, the less number of important attributes included into classification process, the better the performance of Naïve Bayes. The result is on the contrary to SVM. More specifically, the use of NB-CFS on Biopsy/Cytology test or NB-RF on Hinselmann/Schiller test is acknowledged to build cervical cancer predictive model. Furthermore, Naïve Bayes/ Random Forest is able to produce the most risk factors to cervical cancer and also reduce the computational time of the classification methods.

References
[1] Parkin D M, Bray F I Devesa S S 2001 Cancer burden in the year 2000. The global picture *Eur J Cancer* 37 S4
[2] Irabor G N, Isiwele E, Nnoli M A and Omoruyi K A 2018 The relationship between Age and Histological Types of Cervical Cancer *Int. J. Of Scientific Research* 7(2) 260
[3] Wu W and Zhou H 2017 Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches *IEEE Trans. Knowl. Data Int*.
[4] Punjani D N and Atkotiya K H 2017 Cervical cancer prediction using data mining *Int. J. for Research in Applied Science & Eng. Tech. (IJRASET)* 5(12) 1856
[5] Tomar D and Agarwal S 2013 A survey on Data Mining Approaches for Healthcare *Int. J. of Bio-Science and Bio-Tech.* 5(5) 241
[6] Ceylan Z and Pekel E 2017 Comparison of multi-label classification methods for prediagnosis of cervical cancer *Int. J. of Int'l. System and App. in Eng. (IISJAE)* 5(4) 232
[7] Wu W and Zhou H 2017 Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches *IEEE Trans. Knowl. Data Int*.
[8] Aravithan K and Vanitha M 2016 A Comparative study on Prediction of Heart Disease using Cluster and Rank based Approach *Int. J. of Adv. Research in Comp. and Comm. Eng.* 5(2) 421
[9] Peter T J and Somasundaram K 2012 An Empirical Study on Prediction of Heart Disease using Classification Data Mining Techniques *IEEE-Int. Conf. On Adv. In Eng., Science and Management (ICAESM)* 514
[10] Shouman M, Turner T and Stocker R 2012 Applying k-Nearest neighbor in Diagnosing Heart Disease Patients *Int. J. of Inf. and Education Tech.* 2(3) 220
[11] Dheeru, Dua and Efi K T 2017 *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml] Irvine CA: University of California, School of Information and Computer Science
[12] Kemal A 2018 A Study on Test Variable Selection and Balanced Data for Cervical Cancer Disease *Int. Jour. of Information Eng. and Electronic Business (IJIEEB)* 10(5) 1
[13] Shuib N L, Bakar A A and Othman Z A 2009 Performance Study on Data Discretization Techniques Using Nutrition Dataset *Int. Symp. on Comp., Comm., and Control (ISCCC)* 1 304
[14] HPV-Associated Cancer Diagnosis by Age 2016 *Centers for Disease Control and Prevention* cdc.gov/cancer/hpv/statistics/age.htm