Classification of Cervical Cancer Using Ant-Miner for Medical Expertise Knowledge Management

Juliana Wahid and Hassan Fouad Abbas Al-Mazini

Universiti Utara Malaysia, Malaysia, {w.juliana@uum.edu.my, hassan_fouad@ahsgs.uum.edu.my}

ABSTRACT

The fourth most frequent cause of cancer death in women is cervical cancer. No sign can be observed in the early stages of the disease. In addition, cervical cancer diagnosis methods used in health centers are time consuming and costly. Data classification has been widely applied in diagnosis cervical cancer for knowledge acquisition. However, none of existing intelligent methods is comprehensible, and they look like a black box to clinicians. In this paper, an ant colony optimization-based classification algorithm, Ant-Miner is applied to analyze the cervical cancer data set. The cervical cancer data set used was obtained from the repository of the University of California, Irvine. The proposed algorithm outperforms the previous approach, support vector machine, in the same domain, in terms of better result of classification accuracy.

Keywords: Cervical cancer, ant-miner, data classification.

I INTRODUCTION

The cervical cancer is the fourth most common cause of death from cancer (World Health Org., Geneva, 2014) among female. Cervical cancer starts when cells lining the cervix, i.e. the lower part of the uterus (womb), begin to grow out of control (American Cancer Society, 2017). The normal cells of the cervix first gradually develop pre-cancerous changes that turn into cancer. These changes can be detected by the screening test namely Pap test (Cronjé, 2004) and other imaging techniques such as diffusion-weighted imaging (DWI) and magnetic resonance imaging (MRI) that can reveal cervical cancer to a certain stage (Exner et al., 2016; McVeigh, Syed, Milosevic, Fyles, & Haider, 2008).

However, in the developing and low-income countries, people have low awareness of routine Pap screening test. In addition, limited medical expertise and the lack of medical equipment contributes to higher number of death caused by cervical cancer.

The use of computer and artificial intelligence has improved the health care system, which led to an increase in the demand for intelligent and the discover knowledge in modern medicine practices (Al-behadili, 2018; Djam, Sc, & Kimbi, 2011) Over the past few years, there were several methods proposed and applied for diagnosing and detecting the occurrence of cervical cancer including principal component analysis (PCA), linear regression (LR), particle swarm optimization (PSO), support vector machine (SVM), Artificial Neural Network (ANN), fuzzy positivistic C-means clustering, genetic algorithm (GA), hierarchical decision approach (HDA), texture analysis (Ambrosiadou, Goulis, & Pappas, 1996; Athinarayanan, Srinath, & Kavitha, 2016; Daniel, Hájek, & Nguyen, 1997; Durkin, 1990; Fernandes, Chicco, Cardoso, & Fernandes, 2018; Khan, Maqbool, Razzaq, Irfan, & Zia, 2008; Kingston, 2001; Prasad, Finkelstein, & Hertz, 1996; Ramdhani & Riana, 2017; Soumya, Sneha, & Arunvinodh, 2016).

The models for all these algorithms are usually incomprehensible, and effectively exploiting intelligent systems requires considerable experience.

This paper aim to apply the ant colony optimization (ACO) based classification i.e. Ant-Miner as data classification rules for extraction of relevant features in cervical cancer diagnosis. The Ant-Miner data classification will be used to analyse the cervical cancer dataset (Fernandes, Cardoso, & Fernandes, 2017) from the repository of University of California at Irvine (UCI). Based on literature review, this is the first time that this approach is applied to this dataset. This approach shows that the Ant-Miner can realize the classification of the cervical cancer with high accuracy.

The paper is organized as follows. The related approaches are reviewed in Section II. Section III focuses on the cervical cancer dataset. Then, in Section IV, the performance of the proposed methods when applied to cervical cancer dataset is discussed. Finally, the conclusion is shown in Section V.

II RELATED APPROACHES

There have been many studies in the diagnosis of cervical cancer and on different types of data by using different algorithms.

Aldian, Purwanti, & Bustomi (2013) propose an automatic classification for normal and abnormal cervical cells with artificial neural networks (ANN) and learning vector quantification (LVQ). The sample data sets are collected which performs the steps in digital image processing like pre-processing, filtering and feature extraction. The input image is stored in ANN and for the classification of cervical cells for detection of cancer the LVQ method is used for calculating the coefficient mean value of the extracted image which is used for classifying the normal and abnormal cell with 90% accuracy result.

Sharma, Kumar Singh, Agrawal, & Madaan (2016) diagnose the cancer stage to help treat cancer patients by categorizing the clinical data set for cervical cancer patients. The first step is to divide the image of the Pap smears using the edge detection to separate the cell nucleus from the background and cytoplasm. Next, the extraction of various features of the cervical Pap images such as elongation, perimeter, and region are produced. After that, the min-max method is used to normalize these features. After the normalization step, K-nearest neighbor (KNN) method is used to classify cancer on the basis of its abnormality.

Wu & Zhou (2017) introduced support vector machine (SVM) as an approach to cervical cancer diagnosis. In terms of classification, SVM can classify the coming data into different categories after training. It is in the training process that the learning model was built by dividing original data into different groups via their labels. Between the groups is a hyperplane constructed by SVM, which helps to predict the label of new data. Two improved SVM methods, namely, support vector machine-principal component analysis (SVM-PCA) and support vector machine-recursive feature elimination (SVM-RFE), were suggested to diagnose pernicious cancer samples. They used the cervical cancer data set from the repository of University of California at Irvine (UCI) represented by four target variables, namely, Cytology, Biopsy, Schiller and Hinselmann, and 32 risk factors. The four targets were classified after diagnosis by the three SVM-based approaches. Result showed good accuracy on the use of SVM. The basic SVM method could classify benign and malignant cancer. The other methods could realise analogous functionality with fewer factors than SVM used.

III CERVICAL CANCER DATA SET

The cervical cancer dataset was collected at Hospital 'Universitario de Caracas' in Caracas (Fernandes et al., 2017). The data is represented by 32 risk factors, including demographic information, patient's habits and historic medical records. These features are shown in Table 1. Meanwhile, there are 4 target variables or labels: Hinselmann, Schiller, Cytology and Biopsy. Hinselmanns test refers to colposcopy using acetic acid. Meanwhile, colposcopy using Lugol iodine includes Schillers test, cytology and biopsy. Some patients did not answer all questions for individual privacy reason and accordingly the dataset needs to be pre-treated to deal with the missing values. Considering that the dataset belongs to imbalanced data, oversampling is applied in the pre-treatment process. After pre-treatment, risk factor 27 and 28 were removed as a consequence of lack of available values. Hence we need to analyse cervical cancer data of 858 patients with 30 features.

Table 1. Attributes of Cervical Cancer Data Set.

No.	**Attributes Name**	**Data Type**
1	Age	Int
2	Number of sexual partners	Bool × Int
3	First sexual intercourse (age)	Bool × Int
4	Number of pregnancies	Bool × Int
5	Smokes	Bool
6	Smokes (years)	Int
7	Smokes (packs/year)	Int
8	Hormonal Contraceptives	Bool
9	Hormonal Contraceptives (years)	Int
10	IUD	Bool
11	IUD (years)	Int
12	STDs	Bool
13	STDs (number)	Int
14	STDs: condylomatosis	Bool
15	STDs: cervical condylomatosis	Bool
16	STDs: vaginal condylomatosis	Bool
17	STDs: vulvo-perineal condylomatosis	Bool
18	STDs: syphilis	Bool
19	STDs: pelvic inflammatory disease	Bool
20	STDs: genital herpes	Bool
21	STDs: molluscum contagiosum	Bool
22	STDs: AIDS	Bool
23	STDs: HIV	Bool
24	STDs: Hepatitis B	Bool
25	STDs: HPV	Bool
26	STDs: Number of diagnosis	Categorical
27	STDs: Time since first diagnosis	Int
28	STDs: Time since last diagnosis	Int
29	Dx: Cancer	Bool
30	Dx: CIN	Bool
31	Dx: HPV	Bool
32	Dx	Bool

IV THE PROPOSED METHOD

The algorithm of the proposed method i.e. Ant-miner is shown in Figure 1. The test cases and the training package are classified by using five-fold cross-validation. Ninety percent of the training data and twenty percent of the test data are used in each fold test. After the pheromone initialisation, numerous bases are created in the repeat loop. The procedure is continued with the pruning base and the pheromone update method. When the ants build the same rule consistently more than once (No_Rule_Converg) or the_number_of_ants equals the_number_of_rules, the loop will stop. In the list of rules, the best rule will be added when the inner loop is completed ‘Repeat-Until.’ As a result, all training cases provided in this rule will be removed from the training package. Pheromone is initialised again. The external loop controls the session responsible for configuring this pheromone. For the ‘Repeat-Until’ loop, a limit more than the number of indeterminate training sessions is called Max_uncovered_cases.

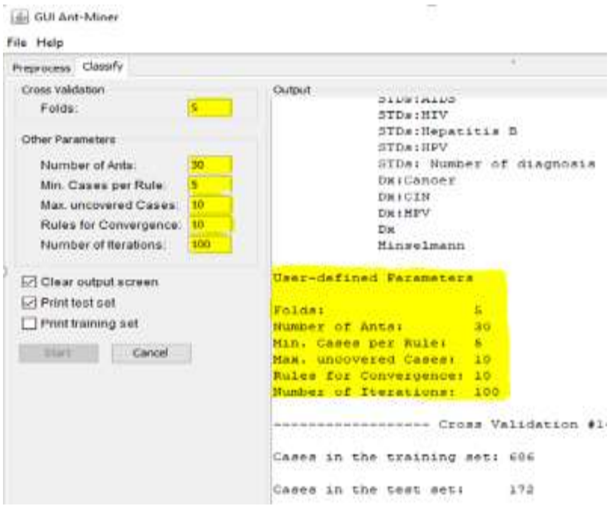
```
Training set = all training cases;
WHILE (No. of uncovered cases in the Training set > max_uncovered_cases)
    i=0;
    REPEAT
        i=i+1;
        Ant_i incrementally constructs a classification rule;
        Prune the just constructed rule;
        Update the pheromone of the trail followed by Ant_i;
    UNTIL (i ≥ No_of_Ants) or (Ant_i constructed the same rule as the previous No_Rules_Converg-1 Ants)
    Select the best rule among all constructed rules;
    Remove the cases correctly covered by the selected rule from the training set;
END WHILE
```

Figure 1. Overview of the Ant-Miner Algorithm.

V EXPERIMENTS AND ANALYSIS

The proposed method is programmed in Java Eclipse and run on a computer Intel (R) Core TM i5 Duo CPU @ 2.40 processor and Windows 10. The parameters used in the proposed method as shown in Figure 2, are Cross Validation = 5, Number of Ants = 30, Min. Cases per Rule = 5, Max. uncovered Cases = 10, Rules for Convergence = 10 and Number of Iterations = 100.

In order to evaluate the performance of the proposed method, the datasets are classified into training and test groups, depending on the number of folds in the cross validation i.e. 5. Four target variables of diagnostic procedures known as Hinselmann, Schiller, Cytology and Biopsy (Jordan et al., 2008), will be diagnosed respectively. The performance evaluation is based on the accuracy rate, rules number, and condition number as shown in Figure 3.



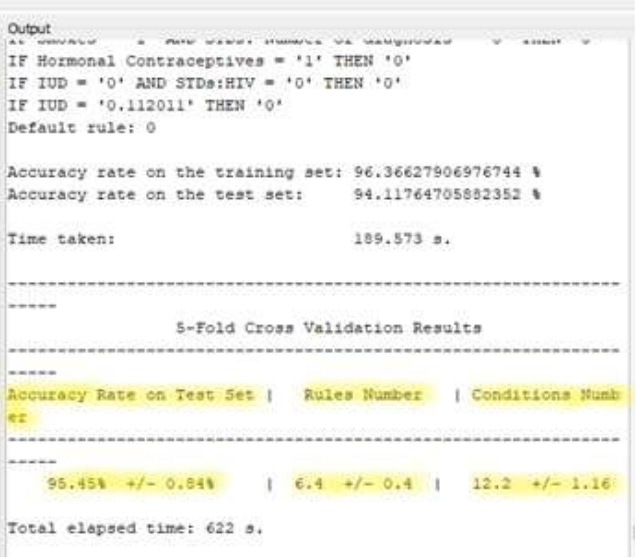
Cross Validation
Folds: 5

Other Parameters
Number of Ants: 30
Min. Cases per Rule: 5
Max. uncovered Cases: 10
Rules for Convergence: 10
Number of Iterations: 100

☑ Clear output screen
☑ Print test set
☐ Print training set

```
                         STDs:AIDS
                         STDs:HIV
                         STDs:Hepatitis B
                         STDs:HPV
                         STDs: Number of diagnosis
                         Dx:Cancer
                         Dx:CIN
                         Dx:HPV
                         Dx
                         Hinselmann

User-defined Parameters

Folds:                    5
Number of Ants:           30
Min. Cases per Rule:      5
Max. uncovered Cases:     10
Rules for Convergence:    10
Number of Iterations:     100

------------------ Cross Validation #1

Cases in the training set: 606

Cases in the test set:     172
```

Figure 2. Parameters Used in the Ant-Miner Algorithm.



```
IF Hormonal Contraceptives = '1' THEN '0'
IF IUD = '0' AND STDs:HIV = '0' THEN '0'
IF IUD = '0.112011' THEN '0'
Default rule: 0

Accuracy rate on the training set: 96.36627906976744 %
Accuracy rate on the test set:     94.11764705882352 %

Time taken:                        189.573 s.

---------------------------------------------------------------
-----
                  5-Fold Cross Validation Results
---------------------------------------------------------------
-----
Accuracy Rate on Test Set |   Rules Number   | Conditions Numb
er
---------------------------------------------------------------
-----
    95.45% +/- 0.84%      |  6.4 +/- 0.4  |   12.2 +/- 1.16

Total elapsed time: 622 s.
```

Figure 3. Results Extracted from the Ant-Miner Algorithm.

The accuracy rate is calculate based on the below Eq. (1):

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

TP is true positive, referring to those malignant cancer samples which have been diagnosed correctly. TN means true negative, equalling to the number of uninfected people who get negative predictions. FP, false positive, is the number of samples without cervical cancer but have been classified into the positive group. Contrary to FP, FN is the number of undetected malignant cancer samples.

The experiment results are shown in Table 2. A total of 30 malignant samples and 638 benign samples are determined in Hinselmann's test. The overall accuracy of the Ant-Miner classification algorithm from 30 features is 95.45%, the rule number is 6.4, and the number of terms per rule is 12.2. Concerning Schiller's test, the number of malignant specimen is 63. The ACO classifier achieves 90.56% accuracy with Schiller's test, the rule number is 6.4, and the number of terms per rule is 14. The Cytology test shows 38 malignant specimens. In addition to perfect diagnosis indexes, the accuracy is 94.64%, with 7.4 for rule number and 14.8 for the number of terms per rule. Unlike the three previous tests, the Biopsy test leads to various detection results. Forty five malignant samples and 623 benign samples are determined. The accuracy is 94.76%, the rule number is 6, and the number of terms per rule is 17.

Table 2. Average and Standard Deviation of Classification Accuracy, Rule Number and Number of Terms Per Rule using Fivefold Cross Validation for Hinselmann, Schiller, Cytology and Biopsy.

Target	Accuracy (%)	Number of Rules	Number of terms per Rules
Hinselmann	95.45 +/- 0.84	6.4 +/- 0.4	12.2 +/- 1.16
Schiller	90.56 +/- 1.18	6.4 +/- 0.4	14 +/- 0.89
Citology	94.64 +/- 0.76	7.4 +/- 0.4	14.8 +/- 1.07
Biopsy	94.76 +/- 1.11	6 +/- 0.32	17 +/- 1.3

Based on the implementation on four target variables, it is shown that Ant-Miner can detect malignant samples and achieve the classification well.

The accuracy results are compared with other approach, i.e. SVM, SVM-PCA and SVM-RFE which also applied on the same cervical cancer data set as shown in Table 3.

Table 3. Comparison of Ant-Miner Classification Accuracy with SVM, SVM-PCA and SVM-FRE.

	Accuracy (%)			
Target	**Ant-Miner**	**SVM**	**SVM-PCA**	**SVM-RFE**
Hinselmann	**95.45**	93.97	93.79	93.69
Schiller	**90.56**	90.18	90.18	90.18
Citology	**94.64**	92.75	92.46	92.37
Biopsy	**94.76**	94.13	94.03	94.03

The comparison of accuracy results indicate that the performance analysis of the Ant-Miner classification achieves a high percentage compared to SVM, SVM-PCA and SVM-RFE for all four target classes.

With the proposed method, it is hope that medical expertise who not specialists in artificial intelligence or machine learning, may be able to utilize this method in clinical practice. Furthermore, any relevant information produced by this method should be appropriately stored for future use.

For future research, this study can be enhanced to test the validity of using pruning procedure to prune only elitist rules instead of pruning each rule constructed by each ant, as proposed by Al-behadili, Ku-mahamud, & Sagban (2018), which aims to achieve more accurate rules and reduce number of terms per rule.

VI CONCLUSION

In this paper, an ACO-based classification algorithm, Ant-Miner was proposed to analyse cervical cancer data set. The proposed method achieved high accuracy results, i.e. more than 90 percent. When comparing with other approaches, the Ant-Miner have higher accuracy results. Thus, this algorithm is the most suitable to cervical cancer data set classification.

REFERENCES

Al-behadili, H. N. K. (2018). Intelligent Hypothermia Care System using Ant Colony Optimization for Rules Prediction. *Journal of Babylon University*, *26*(2), 47–56.

Al-behadili, H. N. K., Ku-mahamud, K. R., & Sagban, R. (2018). Rule Pruning Techniques in the Ant-Miner Classification Algorithm and Its Variants : A Review. In *IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE2018)* (pp. 1–7). Batu Feringghi, Penang, Malaysia.

Aldian, R. D., Purwanti, E., & Bustomi, M. A. (2013). Artificial Neural Network for Classification of Cervical Cancer. *Applied Computing Based*, 4–7.

Ambrosiadou, B. V., Goulis, D. G., & Pappas, C. (1996). Clinical evaluation of the DIABETES expert system for decision support by multiple regimen insulin dose adjustment. *Computer Methods and Programs in Biomedicine*, *49*(1), 105–115. https://doi.org/10.1016/0169-2607(95)01711-9

American Cancer Society. (2017). About Cervical Cancer.

Athinarayanan, S., Srinath, M. V, & Kavitha, R. (2016). Detection and Classification of Cervical Cancer in Pap Smear Images using EETCM , EEETCM & CFE methods based Texture features and Various Classification Techniques, *2*(5), 533–549. https://doi.org/10.18535/ijecs/v5i7.32

Cronjé, H. S. (2004). Screening for cervical cancer in developing countries. *International Journal of Gynecology and Obstetrics*, *84*(2), 101–108. https://doi.org/10.1016/j.ijgo.2003.09.009

Daniel, M., Hájek, P., & Nguyen, P. H. (1997). CADIAG-2 and MYCIN-like systems. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/S0933-3657(96)00376-4

Djam, X. Y., Sc, M., & Kimbi, Y. H. (2011). Fuzzy Expert System for the Management of Hypertension ., *12*(1), 390–402.

Durkin, J. (1990). Research Review: Application of Expert Systems in the Sciences. *The Ohio Journal of Science*, *90*(5), 171–179.

Exner, M., Kühn, A., Stumpp, P., Höckel, M., Horn, L.-C., Kahn, T., & Brandmaier, P. (2016). Value of diffusion-weighted MRI in diagnosis of uterine cervical cancer: a prospective study evaluating the benefits of DWI compared to conventional MR sequences in a 3T environment. *Acta Radiologica*, *57*(7), 869–877. https://doi.org/10.1177/0284185115602146

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. *In Proc. Iberian Conf. Pattern Recognit. Image Anal.*, 243–250.

Fernandes, K., Chicco, D., Cardoso, J. S., & Fernandes, J. (2018). Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Computer Science*, *2018*(5), 1–20. https://doi.org/10.7717/peerj-cs.154

Jordan, J., Arbyn, M., Martin-Hirsch, P., Schenck, U., Baldauf, J. J., Da Silva, D., … Prendiville, W. (2008). European guidelines for quality assurance in cervical cancer screening: Recommendations for clinical management of abnormal cervical cytology, part 1. *Cytopathology*, *19*(6), 342–354. https://doi.org/10.1111/j.1365-2303.2008.00623.x

Khan, F. S., Maqbool, F., Razzaq, S., Irfan, K., & Zia, T. (2008). The Role of Medical Expert Systems in Pakistan, 280–282.

Kingston, J. (2001). High Performance Knowledge Bases: Four approaches to knowledge acquisition, representation and reasoning for workaround planning. *Expert Systems with Applications*, *21*(4), 181–190. https://doi.org/10.1016/S0957-4174(01)00038-0

McVeigh, P. Z., Syed, A. M., Milosevic, M., Fyles, A., & Haider, M. A. (2008). Diffusion-weighted MRI in cervical cancer. *European Radiology*, *18*(5), 1058–1064. https://doi.org/10.1007/s00330-007-0843-3

Prasad, B. N., Finkelstein, S. M., & Hertz, M. I. (1996). An expert system for diagnosis and therapy in lung transplantation. *Computers in Biology and Medicine*, *26*(6), 477–88.

Ramdhani, Y., & Riana, D. (2017). Hierarchical Decision Approach based on Neural Network and Genetic Algorithm method for single image classification of Pap smear. *2017 Second International Conference on Informatics and Computing (ICIC)*, 1–6. https://doi.org/10.1109/IAC.2017.8280587

Sharma, M., Kumar Singh, S., Agrawal, P., & Madaan, V. (2016). Classification of Clinical Dataset of Cervical Cancer using KNN. *Indian Journal of Science and Technology*, *9*(28). https://doi.org/10.17485/ijst/2016/v9i28/98380

Soumya, M. K., Sneha, K., & Arunvinodh, C. (2016). Cervical Cancer Detection and Classification Using Texture Analysis. *Biomedical & Pharmacology Journal*, *9*(2), 663–671. https://doi.org/10.13005/bpj/988

World Health Org., Geneva, S. (2014). *World Cancer Report*.

Wu, W., & Zhou, H. (2017). Data-driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches. *IEEE Access*, 25189–25195. https://doi.org/10.1109/ACCESS.2017.2763984