Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 4 (2013) 275 – 281

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Intelligent Systems and Control

Determination of Significant Features to Precancerous Cervical Classification

A. Aguilera[a*], M. Palma[a], R. Mata-Toledo[b]

[a]*Center of Analysis, Modelling and Treatment of Data, CAMYTD, FACYT, Universidad de Carabobo, Valencia, Venezuela*
[b]*Computer Science Department, James Madison University, Harrisonburg, Virginia, USA*

Abstract

Feature selection is a process used in the automatic learning that consists in selecting an optimal subset of features of a database to reduce its dimensionality, remove noise and improve performance of a learning algorithm. That is, improve the learning speed, precision prediction (measured by the hit rate) and comprehensibility of the results obtained. The aim of this paper is to apply such techniques of dimensionality reduction on processed image features extracted through textural analysis. The processed images were obtained through the colposcope as part of routine gynaecological examinations. From a practical point of view the authors try to extract patterns from the processed images to classify the existing cervix lesions with diagnostic purposes. The resulting attributes of the image processing were analysed using supervised classification techniques of data mining.

Keywords: Feature Selection; Classification, Cervical Lesio

1. Introduction

Cervical cancer is a common type of cancer that begins in the liningcells of the cervix. Initially this abnormal alteration of the structure of the cell is known as cellular dysplasia and is classified as a cervical intraepithelial lesion of low or high grade. In the latter case these abnormal cells can become cancerous if not

* Correspondingauthor: Ana Aguilera. Tel.: +1-540-2461564 ; fax:+1-540-568-6211.
E-mail address: aaguilef@uc.edu.ve

doi:10.1016/j.aasri.2013.10.041

treated on time. Generally, there are no symptoms associated with cervical cancer, so it is essential that women often perform tests to prevent and detect possible lesions as early as possible. Exfoliative cytology is the most common way of an early diagnosis of this disease. There are other methods such as HPV DNA testing, colposcopy exam, visual inspection with acetic acid (AAVI), and visual inspection with Lugol's iodine (LIVI) [3]. The colposcopic inspection is a medical procedure in which a colposcopic camera is used to visually examine the cervix and capture digital images of the same. Often during this test, it is common to use acetic acid or Lugol's iodine to procure a contrast of the cervix and thus aid in the diagnosis of any lesion. Although there are different reasons why cervical cancer can originate, HPV (human papillomavirus) infection is one of the most common. Lesion classification is performed using the Bethesda system established by the National Cancer Institute (NCA) in 1988. This system classifies morphological premalignant lesions into two categories: low-grade squamous intraepithelial lesions (LGSIL) and high-grade intraepithelial lesion (HGSIL). The former category includes the simplest alteration, the reactive inflammatory lesion, suggestive of HPV infection or condylomatousKoilocytosisatypia. This category also includes the next evolutionary level, Cervical intraepithelial neoplasia (CIN) I or mild dysplasia.The HGSIL includes histological lesions CIN II and CIN III or moderate and severe dysplasia, respectively [8] (Fig 1).

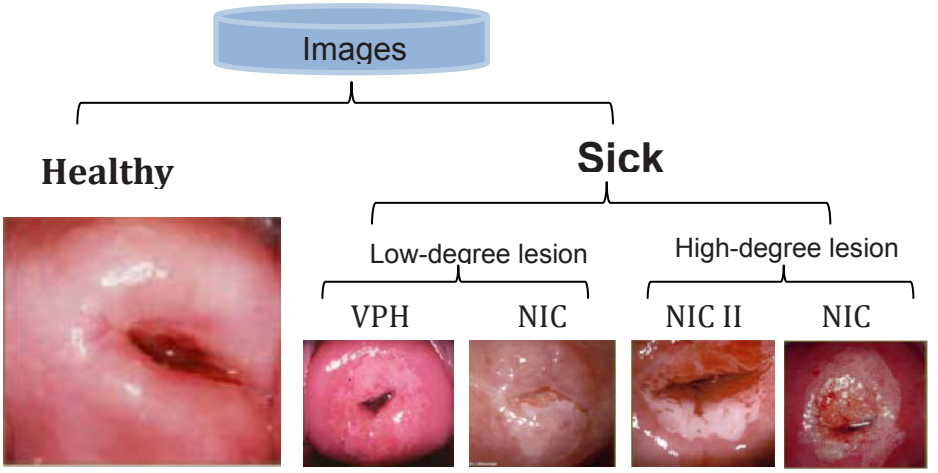

Fig. 1. The classificationof Bethesda system

In recent times there have been major advances from a computational point of view in digital image processing and its subsequent analysis for diagnostic purposes. Parallel to this, techniques such as data mining and machine learning can provide a set of methods that could be used to detect patterns of behavior on large amount of data. One such technique for preparing a database for data mining processing is feature selection. Feature selection serves to identify the best subset of features for a particularly given data mining task. Although the minimum number of attributes that can be used is debatable, for instance, in the classification task, we may assume that the more attributes the higher the discriminatory power. However, several experiments with learning algorithms have shown that it is not always so because, as it has been detected, some experiments have had high runtimes, others have had very high occurrence of redundant or irrelevant attributes while showing a degradation in their classification power [11]. Different experiments have shown that feature selection decreases the error rate of classifiers. This is so because through this process we try to choose the minimal subset of attributes according to following two criteria: first that the hit rate does not drop significantly, on the contrary, it is desirable that it increases. Second, that the distribution of the resulting class be as similar as possible to the original class distribution when all attributes are taken into account. In this

paper, the authors try to compare different methods of feature selection, based on the accuracy of the learning algorithms, with the objective of selecting the best subset of features that provides an effective classification of images for the diagnosis of cervical precancerous lesions.

1.1. Previous works

Current literature surveys show several works related to feature selection methods focused on search techniques, their applications in classification, comparisons, clustering, introduction of new methods, and combination thereof as indicated in [4], [5]. [6], and [8]. In other medical areas the work of Martin et al. [1] applied feature selection methods available in WEKA [2] to a database containing variables involved in the nutritional status of children aged 6 to 11 years. The purpose of that study was to specify which method determined the factors that contributed the most to nutritional assessment. In another study Blakrishnan [3] tried to find an optimal feature subset of the Pima Indian Diabetes Dataset using Symmetrical Uncertainty Attribute Evaluator and Fast Correlation-Based Filter. Guyon et al [7] studied the problem of selecting a small subset of genes from broad patterns of gene expression data recorded on DNA micro-arrays utilizing Support Vector Machine methods based on Recursive Feature Elimination. The studies just mentioned have the common goal of comparing the performance of attribute selection methods with the results obtained by learning algorithms and thus, determining which method significantly improves the results from different situations, with diversity of information, and high or low dimensionality.

2. Data source description

For this research, the authors used a database of cervical digital images from the hospital Maria Teresa del Toro in Maracay (Venezuela). Previously this data was used in an EVA: Recognition System of Precancerous Lesions in the Cervix [9]. This set of images was also used by [10] for thestudy of Pre-cancerous Cervical VideocolposcopicImage Detection using an Artificial Neural Network. The data under study was obtained by taking two images of the cervix of each patient. The first image was taken after an application of acetic acid and, the second, after an application of Lugol's iodine. Each image was classified by physicians in one of the following three categories: a) Healthy: refers cervical images that do not show any injury or alteration; b) BG: refers to images with a LGSIL c) AG: refers to images with a HGSIL. The letters BG and AG, in Spanish, stand for low and high grade respectively; we will continue using these letter combinations throughout the remainder of this paper. The characterization of the images is based on the statistical texture analysis of first and second-order. The first-order is based on the histogram of each plane and the second-order in the co-occurrence matrix. All images were analyzed in RGB (Red, Green, and Blue) planes (Fig 2). To obtain greater accuracy of the learning algorithms, the images were treated in the two following manners: Initially, they were classified using the healthy, BG, and AG categories (Fig. 3). Second, the BG and AG images were grouped in a single sick category. After grouping the images in these categories, the authors ran the learning algorithms to discriminate the images into two groups: healthy and sick. The algorithms were then run anew on the sick category to differentiate between the AG and BG images.

3. Feature selection algorithms

Table 1 shows the algorithms used for feature selection using the data mining tool Weka version 3.6 [2]. The algorithms that evaluate subsets of attributes are distinguished with the letter **s**. Likewise, the algorithms that evaluate the total set of attributes are distinguished with the letter **t**. The **s**-algorithms were combined with

search methods, as show in the table 2, with the exception of the Ranker search method that was used solely the **t**-algorithms.

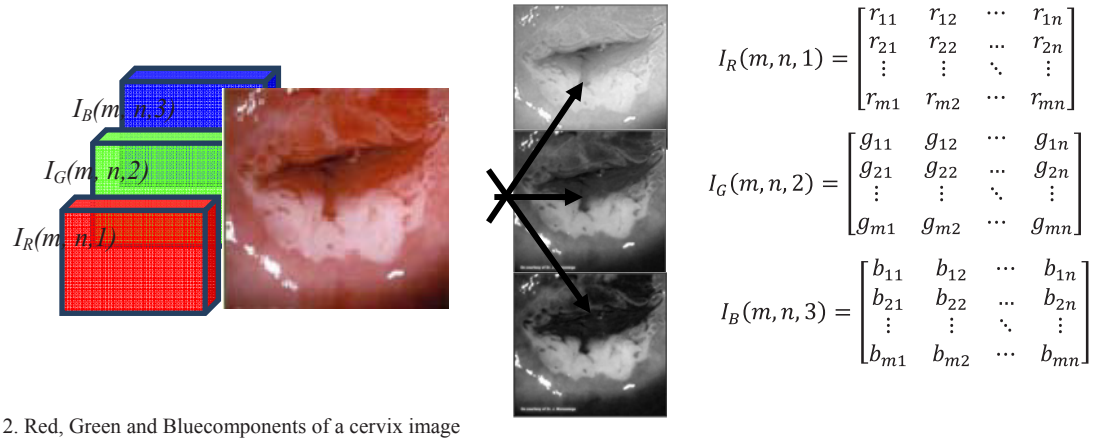

Fig 2. Red, Green and Bluecomponents of a cervix image

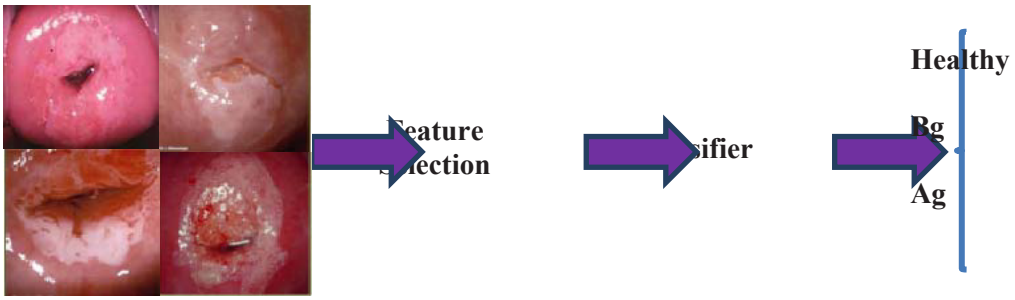

Fig 3. Image processing discrimination in three classes

Table 1. Selection features algorithms

Algorithm	Description
[s]CfsSubsetEval (CFS Correlation–based Feature Selection)	Selects subsets of attributes which have a high correlation with the class and low correlation between themselves
[s]ConsistencySubsetEval	Evaluates the value of the attribute subset according to the level of consistency of the class values when the bodies are projected in the training subset of attributes. The consistency of any subset can never be less than the entire set of attributes
[t]ChiSquaredAttributeEval	Evaluates the value of an attribute by calculating the value of the chi-square statistic with respect to the class
[t]GainRatioAttributeEval	It is a measure of the uncertainty of a random variable based on the concept of entropy from information theory
[t]InfoGainAttributeEval	Finds the set of attributes that provides more information about the class
[t]LatentSemanticAnalysis	Performs latent semantic analysis and transforms data
[t]OneRAttributeEval	Evaluates the value of an attribute using the classifier OneR
[t]PrincipalComponents	Performs a principal components analysis and transforms data
[t]ReliefFAttributeEval	Assigns a weight to each attribute based on the nearest neighbor technique. The weight of each attribute is modified as a function of its ability to distinguish between the values of the class
[t]SVMAttributeEval	Evaluates the value of an attribute by using a classifier Support Vector Machine (SVM). Attributes are classified by the square of the weights assigned by the SVM
[t]SymmetricalUncert [t]AttributeEval	Evaluates the value of an attribute using symmetrical uncertainty with respect to the class

Table 2. Search methods

Search method	Description
BestFirst	Search forward from an empty set using the incremental greedy strategy with backtracking
GeneticSearch	Search using a simple genetic algorithm
LinearForwardSelection	Extensiion of the BestFirst search method
ScatterSearchV1	Performs a dispersed search through the space of the subsets of attributes. Start with a significant population of many and various subsets, and stops when the result is greater than a given threshold or when no further improvement is possible
Ranker	Returns an ordered list of attributes according to their quality, based on their individual assessments.

4. Results and conclusions

The present study was conducted on two groups of cervix images. One group had images with acetic acid application and the other images with Lugol's iodine application. There were 63 textural features extracted from these images; 21 per each layer Red, Green and Blue (R, G and B). Each subset generated by feature selection methods (scenarios in Table 3) was tested with different classifiers.

Table 3 shows the highest percentage of correct answers obtained for each type of image discriminated on the three original classes (Healthy, BG and AG).

Table 3. Best results achieved by each selection method applied to images Lugol's iodine and acetic acid

	Scenario	Imageswith Lugol's Iodine **Classifier**	**%**	Imageswith Acetic Acid **Classifier**	**%**
S1	No Feature selection	DataNearBalancedND RandomForest	73,68	**RotationForest LMT**	71,93
S2	Correlation-based Feature Selection with best first search strategy	RotationForest LADTree	80,70	RandomSubSpace BFTree	63,16
S3	Correlation-based Feature Selection with genetic search strategy	OrdinalClassClassifier RandomForest	75,44	RotationForest LMT	71,93
S4	Correlation-based Feature Selection with linear forward selection search strategy	MultiBoostAB BFTree	75,44	Decorate BFTree	66,67
S5	Correlation-based Feature Selection with scatter search V1 search strategy	**Decorate RandomTree**	82,46	RandomSubSpace BFTree	63,16
S6	Chi-square Feature Evaluation	Decorate RandomTree	77,19	ClassBalancedND FT	64,91
S7	Consistency-based Feature Selection with best first search strategy	RandomForest	77,19	Decorate FT	64,91
S8	Consistency -based Feature Selection with genetic search strategy	Decorate J48	75,44	AdaBoostM1 FT	68,42
S9	Consistency -based Feature Selection with linear forward selection search strategy	Decorate J48	78,95	Decorate J48graft	70,18
S10	Consistency -based Feature Selection with scatter search V1 search strategy	RotationForest RandomTree	80,70	Decorate FT	64,91
S11	Gain Ratio Feature Evaluation	MultiClassClassifier RandomForest	78,95	MultiBoostAB DecisionStump	63,16
S12	Info Gain Feature Evaluation	RotationForest LADTree	77,19	MultiBoostAB DecisionStump	63,16
S13	Latent Semantic Analysis	DataNearBalancedND DecisionStump	63,16	ClassBalancedND FT	56,14
S14	OneR based Feature Evaluation	RandomSubSpace J48	75,44	AdaBoostM1 LMT	70,18
S15	Principal Component Analysis	AdaBoostM1 REPTree	73,68	MultiBoostAB FT	68,42
S16	ReliefF Feature Evaluation	Decorate RandomForest	77,19	Decorate LMT	70,18
S17	SVM based Feature Evaluation	Decorate J48graf	78,95	AdaBoostM1 LMT	68,42
S18	Symmetrical Uncert Feature Evaluation	Decorate RandomTree	80,70	MultiBoostAB DecisionStump	63,16

For the Lugol's iodine images, the best classification accuracy was obtained with the S5 scenario that correctly classified 82.46% of the images using the metaclassifier Decorate from decision tree RandomTree. In acetic acid image group it was observed that the use of feature selection methods had no benefit for the classification process. This is due to the fact that the highest percentage of correctly classified instances (71.93%) was obtained with S1. In this latter scenario no feature selection method was used and all attributes were considered in the classification process. From all the images under consideration we can observe that the set of images with Lugol's iodine provided better accuracy results based on the percentatge of correctly classified instances.

Table 4 shows the results obtained by discriminating between healthy and sick classes with the two types of images. For the Lugol's iodine images, the highest percentage of rated instances was 89.47%. This was obtained by the metaclassifier AdaBoostM1 from REPTree decision tree using S4. For acetic acid images, the highest percentage obtained was 84.21% using the Decorate metaclassifier from J48 decision tree with a S5.We can also observe that the percentage of correctly classified instances is significantly increased when the AG and BG classes were grouped into a single sick class.

Table 4. . Best results achieved by each selection method

	Healthy –Sick				*BG-AG*			
Scenario	Images with Lugol's Iodine	%	Images with Acetic Acid	%	Images with Lugol's Iodine	%	Images with Acetic Acid	%
S1	SMO	84,44	AdaBoostM1 FT	82,22	Decorate RandomTree	87,72	RotationForest REPTree	80,70
S2	Bagging LADTree	84,44	MultiClassClassifier RandomForest	82,22	RotationForest J48graft	85,96	RotationForest DecisionStump	80,70
S3	ClassBalancedND LADTree	86,67	Decorate J48graft	82,22	RotationForest LADTree	85,96	RotationForest REPTree	80,70
S4	Bagging LADTree	84,44	RotationForest J48	84,44	**AdaBoostM1 REPTree**	**89,47**	RotationForest DecisionStump	80,70
S5	Bagging LADTree	84,44	Decorate DecisionStump	82,22	MultiBoostAB DecisionStump	87,72	RotationForest DecisionStump	80,70
S6	**AdaBoostM1 - BFTree**	**86,67**	RandomSubSpace J48	82,22	AdaBoostM1 FT	87,72	MultiBoostAB REPTree	80,70
S7	MultiBoostAB REPTree	84,44	Decorate DecisionStump	82,22	Bagging J48graft	87,72	RotationForest DecisionStump	80,70
S8	Bagging J48	86,67	RotationForest DecisionStump	82,22	MultiBoostAB BFTree	84,21	RotationForest REPTree	80,70
S9	MultiBoostAB REPTree	84,44	**AdaBoostM1 - REPTree**	**84.44**	RandomSubSpace LADTree	89,47	RotationForest DecisionStump	80,70
S10	MultiBoostAB REPTree	84,44	Decorate DecisionStump	84,44	MultiLayerPerceptron	89,47	RotationForest DecisionStump	80,70
S11	AdaBoostM1 - BFTree	86,67	J48	80,00	RotationForest RandomTree	87,72	MultiBoostAB REPTree	80,70
S12	AdaBoostM1 - BFTree	86,67	Bagging RandomForest	82,22	RotationForest RandomTree	89,47	MultiBoostAB REPTree	80,70
S13	Decorate LADTree	80,00	BFTree	71,11	AdaBoostM1 J48	78,95	AdaBoostM1 FT	78,95
S14	RandomSubSpace J48	86,67	MultiLayerPerceptron	84,44	AdaBoostM1 REPTree	85,96	ClassBalancedND RandomTree	80,70
S15	RandomSubSpace BFTree	86,67	RotationForest RandomForest	84,44	AdaBoostM1 LADTree	87,72	**Decorate J48**	**84.21**
S16	RandomSubSpace J48	86,67	MultiLayerPerceptron	84,44	AdaBoostM1 LADTree	85,96	AdaBoostM1 FT	78,95
S17	SMO	84,44	AdaBoostM1 FT	82,22	AdaBoostM1 LADTree	87,72	Decorate RandomTree	84.21
S18	AdaBoostM1 - BFTree	86,67	J48	80,00	AdaBoostM1 FT	87,72	MultiBoostAB REPTree	80,70

The experiments with the sick class for the two types of images produced the best results, namely, the LADTree decision tree provided a 86.67% instances correctly classified in a S3 for Lugol's iodine images and the metaclassifier AdaBoostM1 and REPTree decision tree providing a 89.47% with a S9 for acetic acid images.

Table 5 presents a summary of the scenarios and classifiers that provided the best performance in each experiment, class, and group of images. The best results were obtained by analyzing the cervix images with Lugols' iodine, combining the sick classes, and performing the discriminating classification using only two classes in each case. Currently data analysis real life applications clearly show the need to manipulate a reduced number of attributes. The experiments performed in this study shows that the feature selection is a process that provides significant benefits because the obtained models are more understandable and perform better the learning algorithms than when the complete data set is used.

Table5. Summary of best performance obtained for each set of images and classes.

Classes	Images set	Scenario	Classifier	Accuracy	AbsoluteError
Healthy	With Lugol's Iodine	S5	Decorate - RandomTree	82.4561%	61.2198%
BG-AG	With Acetic Acid	S1	RotationForest - LMT	71.9298%	62.7400%
Healthy	With Lugol's Iodine	S4	AdaBoostM1 - REPTree	89.4737%	45.9349%
Sick	With Acetic Acid	S15	Decorate - J48	84.2105%	121.6727%
BG	With Lugol's Iodine	S6	AdaBoostM1 - BFTree	86.6667%	40.4499%
AG	With Acetic Acid	S9	AdaBoostM1 - REPTree	84.4444%	62.9891%

5. Acknowledgements

The authors appreciate the financial support of Fulbright Program at James Madison University, USA and the CDCH of Carabobo University, Venezuela.

References

[1] Martín, R., Ramos, R., Grau, R. García, M. Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños. GacetaMédicaEspirituana9, 2007.
[2] Weka 3 - Data Mining with Open Source Machine Learning Software. www.cs.waikato.ac.nz/ml/weka/
[3] Balakrishnan, S.Narayanaswamy, R. Feature selection using FCBF in type II diabetes databases. In 7th Annual Conference on Information Science, Technology Management (CISTM), 2009.
[4] Yan, W., Goebel, K. F. Feature selection for partial discharge diagnosis. In 12th SPIE: Health Monitoring and Smart Nondestructive Evaluation of Structural and Biological Systems IV, pp. 166–175, 2005.
[5] Zheng, H. Zhang, Y. Feature selection for high dimensional data in astronomy. Advances in Space Research, 2008.
[6] Forman, G.Guyon, I.Elisseeff, A. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3. pp. 1289–1305, 2003.
[7] Guyon, I., Weston, J., Barnhill, S. Vapnik, V. Gene Selection for cancer classification using support vector machines. Journal of Machine LearningResearch, 46(1-3). pp. 389–422, 2002.
[8] Mejía-Lavalle, M., Solís, J. F., García, F. J. Selección de atributos en una base de datos de facturación eléctrica aplicando programación cóncava. In 4ª Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento, 2004.
[9] Guerrero, J., Pérez, Y. EVA: Sistema de Reconocimiento de Lesiones Precancerosas en el Cuello Uterino. PREGRADO thesis, Universidad de Carabobo., 2005.
[10] Aguilera, A. Guerrero, J. Palma, M.A., Rodríguez, J. Cervical pre-cancerousdetectionfromvideocolposcopicimagesusingan artificial neural network. In Proceedings 4th Indian International Conference on Artificial Intelligence (IICAI), 2009.
[11] Ruiz, R. Heurísticas de selección de atributos para datos de gran dimensionalidad. Ph.D thesis, Universidad de Sevilla, 2006.
[12] Langley, P. Selection of relevant features in machine learning. In The AAAI Fall symposium on relevance, pp. 140–144. AAAI Press, 1994.
[13] John, G., Kohavi, R. Pfleger, R. Irrelevant features and the subset selection problem.pp. 121–129. Morgan Kaufmann, 1994.