Selecting critical features for biomedical data classification

Ulfeta A. Marovac[1*], Lejlija M. Memić[1], Aldina R. Avdić[1], Natasa Z. Djordjević[2], Zana Ć. Dolićanin[3], Goran M. Babić[4]

[1] State University of Novi Pazar, Department of Technical and Technological Sciences, Vuka Karadžića 9, 36300 Novi Pazar, Serbia, e-mail: apljaskovic@np.ac.rs, umarovac@np.ac.rs, lmemic@np.ac.rs
[2] State University of Novi Pazar, Department of Natural and Mathematical Sciences, Vuka Karadžića 9, 36300 Novi Pazar, Serbia, e-mail: natasadj@np.ac.rs
[3] State University of Novi Pazar, Department of Biomedical Sciences, Vuka Karadžića 9, 36300 Novi Pazar, Serbia, e-mail: zdolicanin@np.ac.rs
[4] University of Kragujevac, Faculty of Medical Sciences, Svetozara Markovića 69, 34000 Kragujevac, Serbia, e-mail: ginbabic@medf.kg.ac.rs

* *Corresponding author*

DOI: 10.46793/ICCBI23.136M

Abstract: In this paper, the application of machine learning methods on large data sets with numerous features was investigated, with a focus on the identification of critical features in order to reduce the data and produce more accurate results. The research discusses feature extraction techniques for classifying two biomedical data sets with 62 and 71 features, respectively. The results were compared and presented using four classification techniques. The acquired results demonstrate that the selected important features typically produce more accurate results, or at least the same results while reducing the size of the data set and making data collecting easier.

Keywords: feature selection, machine learning, biomedical data classification, pregnant women

1. Introduction

Prior to applying machine learning techniques to huge datasets with numerous features, it is crucial to prioritize the most significant ones in order to reduce the size of the dataset and produce more accurate results. Choosing good features means choosing only the important features and not leaving out any essential features [1]. High dimension biomedical data usually contains a large number of weak relevant or irrelevant features. Therefore, feature selection is considered to be an essential step in the diagnosis of related diseases using high dimension biomedical data [2]. In classification models for colorectal cancer cases phenotypic [3] as well as for the classification of high-dimensional biomedical data [4], attribute selection produced positive outcomes. This research deals with the problem of determining diagnostic biomarkers in pregnancy. Thrombophilia in pregnancy, preeclampsia and gestational diabetes are a few of the issues that could occur during pregnancy. This study aims to develop a more accurate method of determining whether a pregnant woman has any of the issues above. For this study, biomedical and demographic information was gathered. The research task is to single out the most significant attributes that provide an equally good model for predicting possible problems in pregnancy as a model built on all available attributes. To check the correctness, the classification of pregnant women was applied based on the available biomedical and demographic data using four classification methods and the obtained models were compared with the classification models obtained on a separate set of attributes.

2. Materials and Methods

Subjects were selected from pregnant women who are accommodated at the Gynecology-Obstetrics Clinic of the University Clinical Center Kragujevac for treatment or labor. Four categories of pregnant women were included in the study: 1) healthy pregnant women (control group); 2) women with preeclampsia; 3) pregnant women with thrombophilia; 4) women with gestational diabetes, controlled by diet. Ethics Committee of the University Clinical Center Kragujevac approved the study protocol, and all patients gave informed consent. For the purposes of this research, two sets of data were selected. The first data set contains information on demographic data for 65 women, of which 27 women have biomedical data. These women are from all four categories. The second set of data consists of demographic and biomedical data of 35 who belong to the group of healthy pregnant women or pregnant women with thrombophilia. Biomedical data consists of clinical data obtained by taking an anamnesis, blood and urine analysis, and via ultrasound examination of pregnant women. Demographic data were collected using a questionnaire on demographic and lifestyle issues. The aim of this study is to enhance the model and categorize pregnant women into one of the four or one of the two groups found in the initial data set. Before applying the modeling, we completed identifying the associated features from a collection of data and removing the irrelevant or less significant features. The SelectKBest class (from the Python scikit-learn library) is used to choose 10 of the best features. The chi-squared (chi2) statistical test was performed to determine the most important categorical features (demographic) and ANOVA for numerical data (biomedical analysis). This study compares the performance of K-nearest neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM), and Naive Bayes (NB) machine learning methods for classification on the described datasets. Comparisons were made between the classification outcomes obtained using the set of attributes chosen through univariate selection and the outcomes of applying the methods to all available attributes. The implementation of the proposed machine learning methods is done using Python and corresponding ML packages (pandas, numpy). To compare classification results, the accuracy measure was used.

3. Results and discussion

The proposed method for the selection of attributes for the classification of biomedical data sets was applied to both described sets. The extracted demographic and biomedical features for the two datasets are displayed in Figure 1.

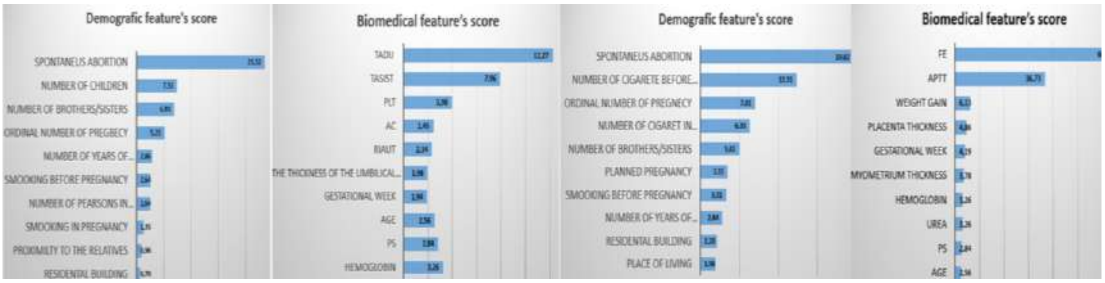

Figure 1. Demographic and biomedical feature's score for the first (left) and the second (right) data set.

A classification using all demographic (biomedical) variables was carried out and the results were compared with those when the classification was carried out just on the basis of 10 selected ones. This was done in order to assess the significance of the selected features on both sets of data. The proposed strategy was evaluated using five-field cross-validation. The first set of data contains four groups of pregnant women, and the classification of this set was made into 4 classes. Four classification methods were applied, and the obtained results are shown in Table 1.

Table 1. Results of 4-class classification on the first data set when using demographic (left) and biomedical (right) data with all attributes and reduced data.

Demographic features					**Biomedical features**				
#Features	**KNN**	**RF**	**SVM**	**NB**	**#Features**	**KNN**	**RF**	**SVM**	**NB**
Top 10	**0.39**	**0.39**	**0.39**	**0.30**	**Top 10**	**0.7**	**0.63**	**0.51**	**0.58**
All (25)	0.39	0.29	0.37	0.25	**All (37)**	0.37	0.45	0.44	0.45

The best classification results for both demographic and biomedical features are obtained using the top 10 extracted features, using the KNN method: 0.39 for demographic and 0.7 for biomedical (Table 1). For all four methods, attribute reduction gave better or at least the same accuracy. The low precision is due to the large number of classes and the small sample. The second set of data shows the results of applying this method when we have only two classes of healthy pregnant women and pregnant women with thrombophilia. Table 2 shows the results obtained by applying the classification to all attributes compared to the classification only to the selected ones. As for demographic attributes, the RF method applied to the selected 10 attributes gave the best results. Again, in all cases, the classification is of the same or better accuracy compared to the reduced attributes.

Table 2. Results of binary classification on the second data set when using demographic (left) and biomedical (right) data with all attributes and reduced data.

Demographic features					**Biomedical features**				
#Features	**KNN**	**RF**	**SVM**	**NB**	**#Features**	**KNN**	**RF**	**SVM**	**NB**
Top 10	**0.80**	**0.83**	**0.80**	**0.63**	**Top 10**	**0.83**	0.91	**0.63**	**1.00**
All (34)	0.66	0.83	0.63	0.57	**All (37)**	0.46	**0.94**	**0.63**	0.97

The classification using biomedical attributes is shown in Table 2, where the best results were given by the NB method using 10 selected attributes. Only in the case of the RF method, less precise results were obtained on the reduced data set compared to the whole set. In all other and global classification methods when applied over 10 attributes give the same or better precision. These results justify the application of the attribute extraction method before creating the classification model. The selection of attributes that are correlated with one disease gave much better results.

4. Conclusions

Based on the presented results, it can be concluded that the application of attribute extraction on biomedical data improves classification results. It has been shown that binary classification gives better results and that, in general, it is necessary to expand the data set. Future studies will focus on data preparation and training set enlargement in order to obtain more precise results.

References

[1] Y. Lyu, Y. Feng, K. Sakurai., *A survey on feature selection techniques based on filtering methods for cyber attack detection*, Information, 14.3 (2023) 191.

[2] J. Tao, Y. Kang., *Features importance analysis for emotional speech classification*, Lecture Notes Comput Sci, 3784 (2015) 449–57.

[3] T. W. Cenggoro, B. Mahesworo, A. Budiarto, J. Baurley, T. Suparyanto, B.Pardamean *Features importance in classification models for colorectal cancer cases phenotype in Indonesia*, Procedia Comput Science, 157 (2019) 313–320.

[4] B. Zhang, C. Peng., Classification of high dimensional biomedical data based on feature selection using redundant removal, PloS one, 14.4 (2019) e0214406.