Predictor Selection for Bacterial Vaginosis Diagnosis Using Decision Tree and Relief Algorithms

Subjects: Artificial Intelligence & Robotics
Created by: Jesus Francisco Perez-gomez, Juana Canul-reich

Requiring only a few relevant characteristics from patients when diagnosing bacterial vaginosis is highly useful for physicians as it makes it less time consuming to collect these data. This would result in having a dataset of patients that can be more accurately diagnosed using only a subset of informative or relevant features in contrast to using the entire set of features. As such, this is a feature selection (FS) problem. In this work, decision tree and Relief algorithms were used as feature selectors. Experiments were conducted on a real dataset for bacterial vaginosis with 396 instances and 252 features/attributes. The dataset was obtained from universities located in Baltimore and Atlanta. The FS algorithms utilized feature rankings, from which the top fifteen features formed a new dataset that was used as input for both support vector machine (SVM) and logistic regression (LR) algorithms for classification. For performance evaluation, averages of 30 runs of 10-fold cross-validation were reported, along with balanced accuracy, sensitivity, and specificity as performance measures. A performance comparison of the results was made between using the total number of features against using the top fifteen. These results found similar attributes from our rankings compared to those reported in the literature. This study is part of ongoing research that is investigating a range of feature selection and classification methods.

Bacterial vaginosis (BV) is a disease affecting millions of women around the world and involves several serious health conditions. It is the most common of the vaginal diseases in women of reproductive age and it is associated with preterm delivery, chorioamnionitis, post-abortion infection, pelvic inflammatory disease, and sexually transmitted diseases, such as human papillomavirus (HPV). This disease can be detected by two clinical procedures: the Amsel criteria and the Nugent score. Another procedure to detect VB is named real-time or quantitative polymerase chain reaction (qPCR), which consists of the extraction, isolation, and amplification of DNA microorganisms present in the vaginal tract. Some of these procedures usually take a long time to analyze the samples and some are invasive. In diseases like bacterial vaginosis, time is a determining factor for its treatment; therefore, requiring fewer data to diagnose BV is useful for physicians. As such, if there was a more effective and less invasive way to detect this disease, the detection and treatment of BV would be faster and more efficient. According to Liang et al., to formulate a diagnosis, physicians generally ask questions related to the symptoms. From this initial small feature set, the physician forms a differential diagnosis and decides what features (questions, exams, laboratory testing, and/or imaging studies) to obtain to rule out diagnoses in the differential diagnosis set. Through hypothetic-deductive reasoning, the most useful features are identified such that when the probability of one of the diagnoses reaches a level of acceptability, the process is stopped and the diagnosis is accepted. It may be possible to achieve an acceptable level of certainty of the diagnosis with only a few features and thus not need to process the entire feature set. Therefore, the physician can be considered a classifier of sorts. Machine learning algorithms (MLAs) based on artificial intelligence (AI) methods can also do this. From this approach, the objectives and motivations of this work were established: first, using MLAs to explore the most relevant attributes in the data of BV that minimizes the needed data to diagnose it, and second, to explore the features or feature subsets that positively impact the performance of the classification algorithms (CAs) regarding diagnosing BV. For this reason, this work was addressed as a feature selection problem. Feature selection (FS) involves the reduction of the number of features (attributes, variables, or predictors) for each instance in a dataset, discarding the less important attributes. The main objectives of FS involve improving the prediction performance of the classifiers, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The support vector machine (SVM) and logistic regression (LR) classifiers were used in this study to measure the ability to classify patients either as positive or negative cases of BV by obtaining the balanced accuracy, sensitivity, and specificity performance...
measures. The dataset used for this objective was obtained from universities located in Baltimore and Atlanta. It contained clinical and biological information about vaginal microorganisms. Finally, the contribution made by this research can be summarized as follows:

(1) The determination of the best fifteen predictors for bacterial vaginosis diagnosis using feature selection algorithms.

(2) Comparison of the results obtained in this research to those obtained in Beck and Foster [2].

(3) The determination of a highly promising combination of SVM as a classification algorithm and decision trees as a feature selector for bacterial vaginosis diagnosis.

References

1. Sophia M. R. Lannon; Kristina M. Adams Waldorf; Tina Fiedler; Raj P. Kapur; Kathy Agnew; Lakshmi Rajagopal; Michael G. Gravett; David N. Fredricks; Parallel detection of lactobacillus and bacterial vaginosis-associated bacterial DNA in the chorioamnion and vagina of pregnant women at term. The Journal of Maternal-Fetal & Neonatal Medicine 2018, 32, 2702-2710, 10.1080/14767058.2018.1446208.

2. Ashley Jones; Bacterial Vaginosis: A Review of Treatment, Recurrence, and Disparities. The Journal for Nurse Practitioners 2019, 15, 420-423, 10.1016/j.nurpra.2019.03.010.

3. David Hilbert; William L. Smith; Sean G. Chadwick; Geoffrey Toner; Eli Mordechai; Martin E. Adelson; Tina J. Aguin; Jack D. Sobel; S. E. Gygax; Development and Validation of a Highly Accurate Quantitative Real-Time PCR Assay for Diagnosis of Bacterial Vaginosis. Journal of Clinical Microbiology 2016, 54, 1017-1024, 10.1128/JCM.03104-15.

4. Huiying Liang; Brian Y. Tsui; Hao Ni; Carolina C. S. Valentim; Sally Baxter; Guangjian Liu; Wenjia Cai; Daniel S. Kermany; Xin Sun; Jiancong Chen; et al. Liya Hejie ZhuPin TianHua ShaoLianghong ZhengRui HouSierra HewettGen LiPing LiangXuan ZangZhiqi ZhangLiyang PanHuimin CaiRujuan LingShuhua LiYongwang CuiShusheng TangHong YeXiaoyan HuangWaner HeWenqing LiangQing ZhangJianmin JiangWei Yujianqun GaoWanxing OuYingmin DengQiaozhen HouBei WangCuichan YaoYan LiangShu ZhangYaou DuanRunze ZhangSarah GibsonCharlotte L. ZhangOulan LiEdward D. ZhangGabriel KarinNathan NguyenXiaokang WuCindy WenJie XuWenqin XuBochu WangWinston WangJing LiBlanca PizzatoCaroline BaoDaoman XiangWanting HeSuiqin HeYugui ZhouWeldon HawMichael GoldbaumAdriana TremouletChun-Nan HsuHannah CarterLong ZhuKang ZhangHuimin Xia Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature Medicine 2019, 25, 433-438, 10.1038/s41591-018-0335-9.

5. Max Bramer; Principles of Data Mining. Undergraduate Topics in Computer Science 2015, 180, 0, 10.1007/978-1-4471-7307-6.

6. Guyon, Isabelle; Elisseeff, André; An introduction to variable and feature selection. JMLR 2003, 3, 1157-1182, 10.1162/153244303322753616.

7. Daniel Beck; James A. Foster; Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData Mining 2015, 8, 23, 10.1186/s13040-015-0055-3.

Keywords

feature selection; bacterial vaginosis diagnosis; microbial communities; SVM; logistic regression

Retrieved from https://encyclopedia.pub/1359