HHS Public Access

Author manuscript

Proc Int Conf Mach Learn Cybern. Author manuscript; available in PMC 2015 April 23.

Published in final edited form as:

Proc Int Conf Mach Learn Cybern. 2014 July 3; 2014: 241–246. doi:10.1109/ICMLC.2014.7009123.

APPLYING MACHINE LEARNING TECHNIQUES IN DETECTING BACTERIAL VAGINOSIS

Yolanda S. Baker[*], **Rajeev Agrawal**[*], **James A. Foster**[†], **Daniel Beck**[†], and **Gerry Dozier**[‡]

Yolanda S. Baker: ysbaker@aggies.ncat.edu; Rajeev Agrawal: ragrawal@ncat.edu; James A. Foster: foster@uidaho.edu; Daniel Beck: danlbek@gmail.com; Gerry Dozier: gvdozier@ncat.edu

[*]Department of Computer Systems Technology, North Carolina Agricultural and Technical State University, Greensboro, NC, USA

[‡]Department of Computer Science, North Carolina Agricultural and Technical State University, Greensboro, NC, USA

[†]Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA

Abstract

There are several diseases which arise because of changes in the microbial communities in the body. Scientists continue to conduct research in a quest to find the catalysts that provoke these changes in the naturally occurring microbiota. Bacterial Vaginosis (BV) is a disease that fits the above criteria. BV afflicts approximately 29% of women in child bearing age. Unfortunately, its causes are unknown. This paper seeks to uncover the most important features for diagnosis and in turn employ classification algorithms on those features. In order to fulfill our purpose, we conducted two experiments on the data. We isolated the clinical and medical features from the full set of raw data, we compared the accuracy, precision, recall and F-measure and time elapsed for each feature selection and classification grouping. We noticed that classification results were as good or better after performing feature selection although there was a wide range in the number of features produced from the feature selection process. After comparing the experiments, the algorithms performed best on the medical dataset.

Keywords

Bacterial Vaginosis; Machine learning; Feature selection; Classification

1. Introduction

Machine learning (ML) utilizes a variety of artificial intelligence and statistical tools to train on past data in order to create reasonable generalizations, discover patterns, classify previously unseen data or predict new directions [1]. The primary objective of ML is to minimize classification errors on the training data. It has the ability to deliver precise or nearly perfect predictions [2]. ML works extremely well on massive datasets that may go beyond the bounds of human analyzation and interpretation. Its utilization runs the gamut and has been applied to many different types of data including leaf specimens, bankruptcy prediction, facial recognition, internet advertisements and a host of other applications. New ML algorithms are being developed and computers are becoming more powerful, which can

lend itself to addressing complex problems with more accuracy and expeditiousness in a way that is practically impossible for humans.

The medical field is quickly embracing machine learning methodologies as these approaches have shown progress in their usefulness in prediction and classification. This implementation could prove useful in discovering ways to lower the cost of medication, improve clinical studies and help facilitate better assessments by physicians [3]. ML can improve the healthcare process as data continues to increase at the same time decreasing the human effort that would traditionally be required. ML has been used in the medical field to diagnose lung cancer, breast cancer, asthma, heart disease, dementia and other diseases and conditions.

There is a minimal amount of published research using supervised machine learning to diagnose BV. In the past few years and as recent as this year, Srinivasan et al. [4], Ravel et al. [5] and Beck & Foster [6] have used both supervised and unsupervised machine learning techniques to classify BV related microbiota. However, we are expanding this research by conducting experiments using a different dataset.

In this paper we use a myriad of feature selection and classification algorithms to identify Bacterial Vaginosis (BV) in women. BV is a very common condition that is signified by changes in vaginal microbiota or microflora. The rest of this paper is organized as follows. Section II features related work in the areas of Bacterial Vaginosis and machine learning. Section III provides details about the feature selection, search method and classification algorithms used for this research. Section IV describes the experiments conducted, metrics and the results. Finally, Section V will present the conclusion and future work.

2. Related Work

BV is often characterized by changes in the vaginal microflora, unfortunately, the causes of those changes are not well understood. Fortunately, it is easily treatable with antibiotics such as metronidazole and clindamycin. BV is most often diagnosed by testing the vaginal fluid via Gram stain and/or by an assessment based on Amsel's clinical criteria. The Gram stain produces a Nugent Score ranging from 1 – 10. A score of seven or greater yields a positive BV diagnosis. On the other hand, three of the following four Amsel's criteria must be present for a positive diagnosis: 1) presence of a fishy like odor, 2) presence of a white discharge, 3) a vaginal pH of > 4.5 and 4) a minimum of 20% "clue cells" detection. However, Nugent's criterion has become the gold standard for diagnosis. In many instances, a diagnosis is made with Amsel's clinical and confirmed with Gram stain. One of the problems women face is that they may be asymptomatic, however, BV positive. BV can cause unfavorable outcomes for women including an odorous discharge, pelvic inflammatory disease (PID), premature labor and cause them to be more susceptible to contracting HIV and other sexually transmitted diseases (STD). The rate at which BV reoccurs is very high and also not well understood [4].

In the world of medicine, machine learning (ML) has been used in the process of simplifying diagnoses and minimizing misdiagnoses. However, it must be noted that this technology is a tool and does not replace the role of the physician; instead, it should be used

to aid in the overall diagnostic process and evaluation of patients. Computer scientists' use of ML techniques on medical data is continuing to rise as they look for patterns to assist with diagnoses and enhancement of patient care [7]. As we see improvements and the generation of new ML algorithms, we will see a decrease in the time it takes to diagnose and an increase in precision, effectiveness and satisfied patients. ML algorithms have gained a much deserved reputation in research for use in assisting with the diagnoses of numerous diseases [8]..

For example, cancer is the second leading cause of death worldwide with lung cancer as the number one cause of cancer death. The American Cancer Society has predicted that in 2014 there will be 224,201 new cases of lung cancer and it will claim 159,260 lives in the Unites States alone [9]. Lung cancer like most cancers can begin to progressively spread to other organs if it goes untreated, and even then, there's a possibility that the treatment may not work. In order to increase the likelihood of eradicating the cancer and increasing the survival rate, early detection and treatment is quintessential. Unfortunately, some of the current testing methods such as Computed Tomography (CT) scan, chest radiography and Sputum analysis either require an extensive amount of time, money and/or can only detect the cancer in its advanced stage, thus, lowering the chances of survival [10].

Machine learning (ML) not only finds its place in the field of medicine, but has also been very beneficial in other applications. Algorithm performance is often highlighted as an ML outcome, and should be, but there are others that should also be taken into consideration such as increase in quality of life, lives saved, interventions implemented and time, effort and money conserved to name a few. These additional outcomes can help connect ML to other real world problems. It's not enough to simply run an algorithm on dataset, it should include determining the most relevant features, analyzing and interpreting the results and convincing others that this technique is worthwhile for large scale implementation [11].

The field of biometrics has embraced machine learning to assist in the identification and authentication process. There are several modes of biometric identification including fingerprints, iris, signature, voice and face. Shelton et al. [12] developed the Genetic and Evolutionary Feature Extraction – Machine Learning ($GEFE_{ML}$) algorithm for facial recognition in the area of Genetic & Evolutionary Biometrics (GEB). This algorithm works based on the principles of Darwinism's natural selection. They compared the performance of their $GEFE_{ML}$ with that of the traditional Local Binary Pattern (LBP) feature extraction technique. $GEFE_{ML}$ accuracy was comparable to LBP and reduced processing time by 45% (in terms of computational complexity).

3. Feature Selection, Search Method and Classification Algorithms

Feature selection (FS) is the process of choosing the most significant features and forming a subgroup or subset that will be the most valuable for prediction and analysis. The goal is to discover a subset of features that perform as well (or better) than the original set. There is an assumption that datasets include irrelevant, noisy and duplicate data [13]. One of the benefits of ML is feature selection. FS reduces the amount of data that has to be analyzed in turn reducing storage and runtime. This pre-processing step may cost you time in the

beginning, but will improve the outcome and efficiency in the end. This is especially true when dealing with enormous amounts of data. In addition, by executing FS we can anticipate that algorithms will learn more quickly and accuracy will be improved because irrelevant features have been reduced or completely eliminated.

There are two main categories of FS: minimum subset and feature ranking. Minimum subset algorithms produces maximum results with a smaller feature subset and feature ranking merely ranks the features based on specific evaluation measures.

The two primary FS approaches fall within the two categories listed above: filter methods and wrapper methods. Filters create a subset before learning begins that is the most favorable. Based on overall characteristics, an autonomous evaluation is made. Because filters run much faster than wrappers, they may be the preferred method for large and highly dimensional datasets. Wrappers assess the subset by "wrapping around" a classification algorithm that will be used for learning. They usually outperform filters in terms of accuracy; however, the computational cost is very high when used on large datasets. Feature selection algorithms are typically coupled with a search method such as genetic search, exhaustive search and best first [14]. A given search method will roam through the features in order to locate good subsets.

Classification falls under the purview of supervised learning. The objective of a classifier algorithm is to accurately group objects into a predefined set of classes. In other words, it predicts the class of each instance [15]. Thisproach is mostly used in artificial intelligence (AI), machine learning and pattern recognition. Just as with machine learning, classification has been used in a variety of applications such as medical diagnosis, biometrics, cybersecurity, risk analysis, manufacturing, etc. [16]. Choosing the best classifier for a particular problem is extremely important, yet this task has not been given much research attention [17].

The following sub-sections will list and describe the feature selection, search method and classification algorithms that produced the top three results for each set of experiments. All of the algorithms used for these experiments are listed in [18]. In addition to those described below, Weka has an additional six feature selection, seven search method and 89 classification algorithms.

3.1. Feature Selection Algorithm

All top three results in both experiments use the Wrapper Subset Eval feature selection algorithm. It uses a classifier to determine a subset of features. Three of the hosts of available classifiers are OneR, Bagging and NaïveBayes. However, cross-validation is used to approximate the precision of the learning scheme for the feature subset.

3.2. Search Methods

- BestFirst: Explores a random subset of features using greedy hill climbing and supplemented with backtracking. Backtracking is controlled by selecting the number of sequential non-improving nodes allowed. An empty set of features may be initially selected for a forward search, a full feature set for a backward search or

begin midway and search both ways so that all possible distinct feature additions and deletions at any location can be examined.

- Genetic Search: Based on the principles of evolution's survival of the fittest, the genetic search begins with an empty feature set along with rules generated randomly for the initial population. Afterwards, new populations and offspring are formed from the rules of the current population. Crossover and mutation are administered to create offspring. This process repeats until every rule in final population fulfills the fitness threshold.
- Linear Forward Selection: Is an extension of Best First. The user selects m number of features that should not be exceeded in each step. Runtime is reduced because the number of evaluations has been decreased. Linear Forward Selection uses one of two methods; fixed set or fixed width. Both rank the features using a subset evaluator. Fixed set uses only the m best features in the succeeding forward selection while fixed width increases k in each successive step.
- Subset Size Forward Selection: Is an extension of Linear Forward Selection. The search executes k-folds cross validation that can be specified by the user. The prime subset-size is then chosen by executing a Linear Forward Selection on every fold. Lastly, the whole data set is used to execute a Linear Forward Selection up to the prime subset-size.

3.3. Classification Algorithms

- Bagging: Uses a random classifier and combines or aggregates copies of that classifier to improve performance. Bagging for classification takes a majority vote for a predicted class by a sequence of classifiers.
- Random Forest: A collection or ensemble of decision trees. It uses the outcomes of the trees that are individually "weak" classifiers to make one strong classifier. This is done by way of each tree voting on the most common class.
- NaïveBayes: a simple probabilistic classifier based on the supposition of class conditional independence of features and that the prediction is not biased by any concealed features.
- RBF Network: Comprised of three layers: input, hidden and output, it is similar to the k-means algorithm in that the expected target value will most likely have similar values of those that are nearby. The name radial basis function derived its name because it uses radius distance.

4. Experiments and Results

4.1. Dataset

In this section, we provide our experiment process using the machine learning techniques defined in Section 3. The dataset used in our experiment is comprised of 25 women studied over a 10 week period. This data is a subset of a larger dataset of 400 women (Ravel et al., 2011). Dr. James A. Foster and Daniel Beck form the University of Idaho provided us with

the de-identified data in a .csv file. The study was arranged so that samples and information were retrieved from the women every day during the 10 week period, however, some women missed days. There were also a few weeks that are void of any data in the spreadsheet. There are a total of 1601 instances and 418 features. The original BV dataset consisted of three sub-categories of features: time series, clinical and medical data. For this set of experiments, we used only the clinical and medical data.

4.2. Experiments

For all of our experiments, we used the Weka workbench. Weka, written in Java, has a compilation of data preprocessing tools and machine learning algorithms. The experiment process is shown in Figure 1. We used a combinations of the five feature selection, six search methods and three classifier algorithms (used for wrapper methods) assembled to create 20 distinct feature selection sets (FSS). In addition, we selected nine classification algorithms for our experiments. The default settings were maintained for all feature selection (FS), search method and classification (CL) algorithms. We used 10-fold cross-validation for testing and training.

4.2.1. Clinical Experiment Processes—For the clinical data experiment, we retained only the columns containing the clinical data (features 12 – 38) which included questionnaire results and Amsel's clinical criteria. Feature selection and classification algorithms were applied to both giving us information on time elapsed and metric results. Tables were created from this output.

4.2.2. Medical Experiment Processes—In the medical data experiment, we retained only the columns containing the medical data (features 39 – 418) which was derived from the data obtained via the 454 sequencing of the V12 region of the 16S gene. We calculated the time taken for each feature selection and classification algorithm to produce output. We then created an elapsed time table and additionally created feature set and metrics tables.

4.3. Metrics Defined

In classification where there are solely two classes such as with our data where yes = BV positive and no = BV negative, there are only four possible outcomes shown in the confusion matrix in Figure 2.

In the framework of our research, the confusion matrix components have the following descriptions:

- True positive (TP) is the number of correctly classified positive cases of BV,
- False negative (FN) is the number of positive cases of BV incorrectly classified as negative,
- False positive (FP) is the number of negative cases of BV incorrectly classified as positive, and
- True negative (TN) is the number of correctly classified negative cases of BV.

The overall accuracy (AC) is the percentage of correctly classified cases of BV. It is calculated using the number of correctly classified instances, TP and TN divided by the total number of classified BV cases:

$$AC=\frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

The precision (PR) is the percentage of positive predictions retrieved that were actually positive cases of BV. It is the number of true positives divided by the number of all retrieved positive results:

$$PR=\frac{TP}{TP+FP} \quad (2)$$

The recall (RC) is the percentage of positive predictions retrieved from all positive cases of BV. It is the number of true positives divided by the number of all positive cases of BV:

$$RC=\frac{TP}{TP+FN} \quad (3)$$

The F-measure (FM) is the harmonic mean of precision and recall. The harmonic mean is usually used when determining the average of rates:

$$FN=2*\frac{PR*RC}{PR+RC} \quad (4)$$

4.4. Results

In this section we present the results from our research on the clinical and medical only datasets. Both datasets were used in [18], but in this paper, we wanted to see the effects of medical and clinical features independently.

4.4.1. Clinical Results—Table 1 shows the following clinical results:

- Wrapper Sublet Eval / Genetic Search / Bagging feature selection set (FSS) with Bagging classification (WGBB) had slightly better results for precision and accuracy.
- Wrapper Sublet Eval / Best First / Bagging FSS with Random Forest classification (WBBR) and Wrapper Sublet Eval / Subset Forward Selection / NaïveBayes FSS with Random Forest (WSNR) had better results for recall, F-measure and a smaller feature set: 10 compared to 19 for WBBR.
- WSNR had better time than WGBB.
- Based on the results, we have determined that WSNR is the better algorithm for this dataset.

4.4.2. Medical Results

- Table 2 displays the medical results:
- Wrapper Sublet Eval / Best First / Bagging FSS with Bagging classification (WBBB) had slightly better results for accuracy.
- Wrapper Sublet Eval / Linear Forward Selection / NaïveBayes FSS with RBF Network (WLNR) and Wrapper Sublet Eval / Subset Forward Selection / NaïveBayes FSS with RBF Network (WSNN) had better results for precision, recall and F-measure.
- WSNN had better time than WLNR.
- We have determined that WSNN is the better algorithm for this dataset.

5. Conclusion and Future Work

When considering overall accuracy, runtime, reduction in features and recall, we have determined that WLNR of the medical dataset is the better algorithm to use for this problem. We gave more weight to recall than precision because if BV goes undiagnosed and therefore untreated, it can cause very harmful effects for women as there is an increased chance of pre-term labor and pelvic inflammatory disease (PID).

On the other hand, if a woman is diagnosed as BV positive but in reality is negative (false positive); the consequence will merely be taking an inexpensive anti-biotic which will cause little to no harm for women. While the difference between the false negative outcomes for this data seems minimal, the fact that approximately 1 million pregnant women are diagnosed with BV yearly highlights the significance of the results. Our future work will be dedicated towards expanding the algorithm selections, manipulation of the default settings and adjusting seed values for randomization of deterministic algorithms.

Acknowledgements

This research was funded by the National Science Foundation (NSF), Science & Technology Center: Bio/Computational Evolution in Action Consortium (BEACON) and NIH INBRE award P20GM016454. The author would like to thank the NSF, BEACON and NIH for their support of this research.

References

1. Hosseinzadeh F, KayvanJoo AH, Ebrahimi M, Goliaei B. Prediction of lung tumor types based on protein attributes by machine learning algorithms. Springer Plus. 2013; 2(1):1–14. [PubMed: 23419944]
2. Anu, J.; Agrawal, R.; Bhattacharya, S. IEEE Southeast Con. Lexington, KY: 2014. Ranking Tourist Attractions using Time Series GPS Data of Cabs. in-press.
3. Salama GI, Abdelhalim M, Zeid MA-e. Breast Cancer Diagnosis on Three Different Datasets using Multi-classifiers. International Journal of Computer and Information Technology. 2012; 1(1):36–43.
4. Srinivasan S, Marrazzo JM, Fredricks DN, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW, Ross FJ, McCoy CO, Bumgarner R. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PloS one. 2012; 7(6):e37818. [PubMed: 22719852]
5. Ravel J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J. Vaginal microbiome of

reproductive-age women. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(Suppl 1)(11):4680–4687. [PubMed: 20534435]

6. Beck D, Foster JA. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PloS one. 2014; 9(2):e87830. [PubMed: 24498380]
7. Savage N. Better medicine through machine learning. Generic, ACM. 2012:17–19.
8. Filippo A, Alberto L, Eladia Maria P-M, Petr V, Josef H. Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine. 2013; 11(2):47–58.
9. Society AC. What are the key statistics about lung cancer? 2014 Feb 13. http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics.
10. Taher F, Sammouda R. Lung cancer detection by using artificial neural network and fuzzy clustering methods. :295–298.
11. Wagstaff K. Machine learning that matters. arXiv preprint arXiv:1206.4656. 2012
12. Shelton J, Alford A, Small L, Leflore D, Williams J, Adams J, Dozier G, Bryant K, Abegaz T, Ricanek K. Genetic & evolutionary biometrics: feature extraction from a machine learning perspective. :1–7.
13. Hall, MA. Doctoral dissertation, Computer Science. The University of Waikato; 1999. Correlation-based feature selection for machine learning.
14. Witten, IH.; Frank, E.; Hall, MA. Data mining: practical machine learning tools and techniques, third edition. Burlington, Mass: Morgan Kaufmann Publishers; 2011.
15. Dua, S. Data Mining and Machine Learning in Cybersecurity. Boca Raton: CRC Press; 2011.
16. Anbarasi M, Anupriya E, Iyengar NCSN. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology. 2010; 2(10):5370–5376.
17. Peng Y, Kou G, Ergu D, Wu W, Shi Y. An Integrated Feature Selection and Classification Scheme. Studies in Informatics and Control. 2012; 21(3):241–248.
18. Baker, Y.; Agrawal, R.; Foster, JA.; Beck, D.; Dozier, G. Detecting Bacterial Vaginosis Using Machine Learning. 52nd Annual ACM Southeast Conference; Kennewaw, GA. in press.

Figure 1.
Experiment Process

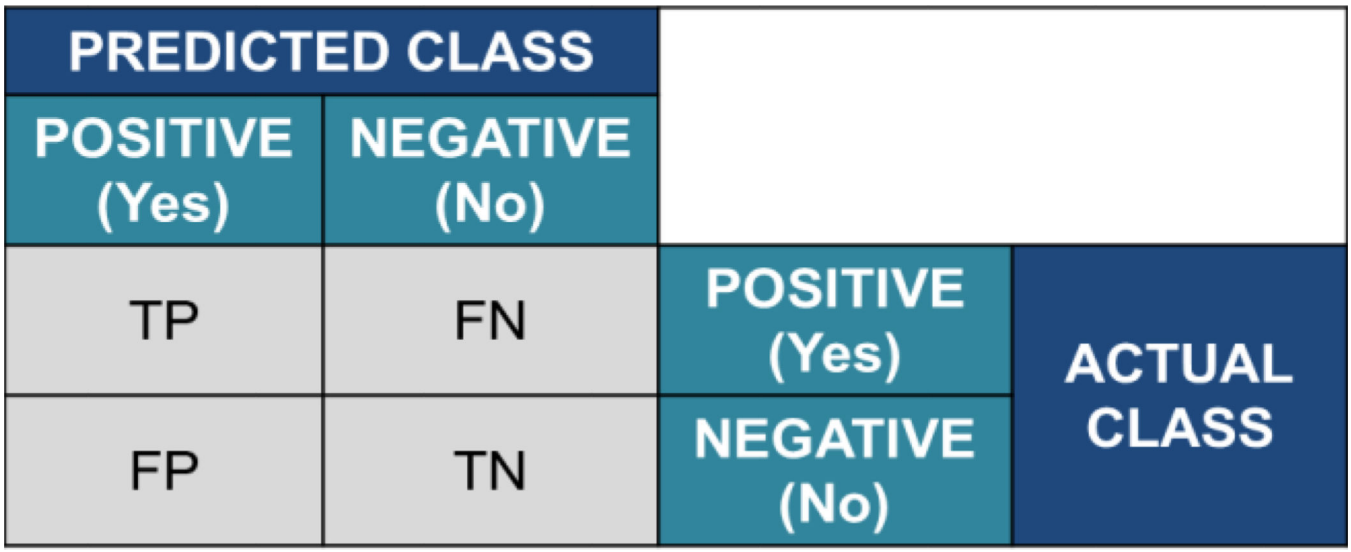

Figure 2.
Confusion Matrix.

Table 1

Top 3: Clinical Results

	WBBR	WGBB	WSNR
Accuracy	89.1318%	89.6315%	89.1318%
Precision	0.714	0.771	0.714
Recall	0.502	0.474	0.502
F-Measure	0.59	0.587	0.59
# of Feat.	10	19	10
FS Time	0:02:30	0:07:19	0:01:34
CL Time	0:00:02	0:00:01	0:00:02

Table 2

Top 3: Medical Results

	WBBB	WLNR	WSNN
Accuracy	94.7533%	95.7527%	95.7527%
Precision	0.857	0.876	0.876
Recall	0.795	0.847	0.847
F-Measure	0.825	0.861	0.861
# of Feat.	14	14	14
FS Time	1:19:37	0:01:32	0:01:07
CL Time	0:00:01	0:00:01	0:00:00