HHS Public Access

Author manuscript

Proc 2014 ACM Southeast Reg Conf. Author manuscript; available in PMC 2016 January 07.

Published in final edited form as:

Proc 2014 ACM Southeast Reg Conf. 2014 March ; 2014: . doi:10.1145/2638404.2638521.

Detecting Bacterial Vaginosis Using Machine Learning

Yolanda S. Baker,

North Carolina A&T State University, 1601 E. Market St, Greensboro, NC 27411,
ysbaker@aggies.ncat.edu

Rajeev Agrawal,

North Carolina A&T State University, 1601 E. Market St, Greensboro, NC 27411,
ragrawal@ncat.edu

James A. Foster,

University of Idaho, 709 S. Deakin St, Moscow, ID 83844, foster@uidaho.edu

Daniel Beck, and

University of Idaho, 709 S Deakin St, Moscow, ID 83844, danlbek@gmail.com

Gerry Dozier

North Carolina A&T State University, 1601 E. Market St, Greensboro, NC 27411,
gvdozier@ncat.edu

Abstract

Bacterial Vaginosis (BV) is the most common of vaginal infections diagnosed among women during the years where they can bear children. Yet, there is very little insight as to how it occurs. There are a vast number of criteria that can be taken into consideration to determine the presence of BV. The purpose of this paper is two-fold; first to discover the most significant features necessary to diagnose the infection, second is to apply various classification algorithms on the selected features. It is observed that certain feature selection algorithms provide only a few features; however, the classification results are as good as using a large number of features.

Keywords

Machine Learning; Feature Selection; Classification

General Terms

Algorithms

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

1. INTRODUCTION

Machine learning (ML) algorithms are centered around prediction based on generalizations made from previous examples. The provision of larger amounts of data allows for tackling larger problems [2]. ML is used in business, academia, government, science and other industries. Its application runs the gamut and has been applied to many different types of data including fraud detection, labor negotiations, facial and iris recognition [9], email messages and many other applications.

ML has also been used in the medical field to diagnose breast cancer [5], asthma [6], dementia [12] and other diseases and conditions. It has recently been used to compare the performance of a variety of classification algorithms in detecting breast cancer [14]. The algorithms compared were Bayes Network, Radial Basis Function (RBF) Networks, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm.

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. In section II we feature 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 and results and finally Section V will present the conclusion and future work.

2. RELATED WORK

As mentioned above, BV is often characterized by changes in the vaginal microbiota. Unfortunately, the causes of those changes are not well understood [10]. BV is most often diagnosed by testing the vaginal fluid via Gram stain and/or by assessing based on Amsel's clinical criteria. The Gram stain produces a Nugent Score ranging from 1 – 10. A score of 7 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 over 4.5 and 4) a minimum of 20% “clue cells” detection [4]. However, Nugent's criterion has become the gold standard for diagnosis [7]. In many instances diagnosis is made with Amsel's clinical and confirmed with Gram stain. A problem women face is that some women may be asymptomatic although BV positive [11]. BV can cause unfavorable outcomes for women including an odorous discharge, pelvic inflammatory disease (PID) and premature labor as well as make them more susceptible to HIV and other sexually transmitted diseases (STD) [2].

Srinivasan et al. [10] performed deep sequencing of the 16S rRNA gene in an attempt to uncover the variety and make-up of vaginal bacteria in BV positive women. They discovered that there were only two bacteria, *Leptotrichia amnionii* and *Eggerthella* sp. that were linked to all four Amsel's criteria. They also uncovered the fact that there was a greater presence of *Lactobacillus crispatus* or *Lactobacillus iners* in women without BV. Unsupervised machine learning (clustering) was one of the methodologies used to make taxonomic connections with BV. They concluded that vaginal bacteria biota in women with

BV is varied and in greater quantities in addition, there was no single bacterium present in 100% of the women.

Beck & Foster [1] applied genetic programming, random forests, and logistic regression machine learning techniques on two BV datasets from Srinivasan et al. [10] and Ravel et al. [8] to hopefully discover BV related microbial relationships. While the associated microbe clusters were different in the two datasets, they did discover that some of the clusters had overlapping microbes. They performed experiments on both the Nugent score and Amsel's criteria. Their experiments resulted in logistic regression and random forest outperforming genetic programming. On the Nugent score logistic regression and random forest maintained accuracy between 90% and 95%. Amsel's criteria produced slightly lower accuracy. However, none of the three classification algorithms fell below 80% accuracy.

There is a minimal amount of published research using supervised machine learning to diagnose BV. However, ML has often been applied in medicine such as in diagnosing and detecting breast cancer. ML has the ability to identify patterns that may be otherwise difficult to detect as well as learn from previous instances. While there has been much research, it is still unknown what causes breast cancer. Due to this fact, early detection is imperative in reducing the death rate. This detection should be able to reliably and accurately differentiate between malignant and benign tumors. The customary technique for diagnosis is usually performed by human observation. However, the number of patients is increasing and therefore using computers to aid and automate the diagnosis process has been developed in the past several years. The qualitative data is converted to a quantitative feature classification problem which has more objectivity [5].

3. FEATURE SELECTION, SEARCH METHOD AND CLASSIFICATION ALGORITHMS

3.1 Feature Selection

Feature selection 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. It is assumed that datasets contain data or features that are not relevant as well as duplicate data. A major benefit of using feature selection is that it reduces the amount of data that has to be analyzed in turn reducing storage and runtime [3]. The following feature selection algorithms in “Table 1” were used for this research.

Feature selection generally falls into one of two categories: minimum subset and feature ranking. Minimum subset algorithms produce exactly what the name suggests; it creates a subset of features with the least amount of relevant features that will yield maximum results. However, there is no distinction between the features in terms ranking. On the other hand, feature ranking algorithms do not reduce the dataset but instead orders the features based on evaluation measures that have been specified.

3.2 Search Methods

A given search method will roam through the features in order to locate good subsets. Each feature selection algorithm is coupled with a search method. We used the search methods as shown below in “Table 2”.

3.3 Classification Algorithms

Classification as previously mentioned, is one of two supervised learning techniques. 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. This approach is mostly used in artificial intelligence (AI), machine learning and pattern recognition. Just as machine learning, classification has been used in a variety of applications such as, medical diagnosis, biometrics, cybersecurity, risk analysis, manufacturing, etc. “Table 3” lists the classification algorithms applied on the reduced feature sets.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

The dataset used in our experiment is comprised of 25 women studied over a 10 week period. The study was setup so that samples and information was retrieved from the women every day during that period; however, some women missed days and even a few weeks. There are a total of 1601 instances and 418 features. The BV data contains three sub-categories: time series, clinical and medical data. The time series data documents day and week numbers (day 1, week 9), day of the week (Monday = 2), day number of the study (1, 2...70) and patient id number (1, 2...25). The clinical data is a combination of results from a questionnaire and Amsel's criteria. The questionnaire included questions such as sexual activity, contraceptive use, tobacco use, etc. The medical data was obtained from vaginal swabs which were used to perform deep sequencing of the 16S rRNA gene coupled with species-level taxonomic identification [10].

4.2 Weka

We used the Weka (Waikato Environment for Knowledge Analysis) workbench to run our experiments. Weka was created at the University of Waikato in New Zealand and has a compilation of data preprocessing tools, data mining and machine learning algorithms. Weka's interface allows users to easily use the tools and apply algorithms on a variety of datasets by accessing the “Explorer” through its graphical user interface (GUI). It was written in Java and runs on most operating systems such as Windows, Mac OS and Linux.

The Weka Explorer makes available the following features for working with data: Preprocessing, classification, clustering, association rule mining, attribute selection and visualization. Weka imports data files that are in the Attribute-Relation File Format (ARFF), comma-separated values file format (CSV) as well as a few others. Once you load your dataset, there are many options to choose from including classification and clustering options such as, cross-validation, percentage split and classes to cluster evaluation.

4.3 Experiment

The experiment process as shown in “Fig. 1” was executed as follows. First we converted the data into a format that was acceptable for use in Weka. We deleted the first column which was row identifier not included as a feature. The Nugent Score results were contained in the second column. If the Nugent score was 7 and 10, it was given a score of “1” indicating BV positive; otherwise, it was given a score of “0” indicating BV negative. We converted all 1's to “yes” and 0's to “no” in the nugentScoreBV column to grant access to a greater number of Weka's algorithms. The data in the nugentScoreBV column was moved to the last column of the dataset as required by Weka for supervised learning.

Weka has two groups of feature selection algorithms: Attribute subset evaluators and single-attribute evaluators. The single-attribute evaluator algorithms only reduce the feature set in theory in that it uses the Ranker method and ranks all of the features in descending order. It requires the user to manually set a threshold to discard the lower ranking features or specify the number of features to preserve. We used five of the six attribute subset evaluator algorithms featured earlier in “Table 1” on the full training set because they automated the feature reduction process.

Weka has a total of 93 classification algorithms 27 of which were not available for the type of dataset used, leaving 66 for classification. We quickly realized that it would be a formidable task to run experiments using all of the available algorithms, so to begin the classification process; we reduced our algorithm selection down to nine popular algorithms. Final classification algorithm selection was shown previously in “Table 3”.

Several of Weka's feature selection algorithms were used for our feature selection process and classifier algorithms were used to determine whether or not we had the ability to accurately diagnose the presence of Bacterial Vaginosis (BV). “Table 4” features the combinations of five feature selection, six search method and three classifier algorithms (used for wrapper methods) assembled to create 20 distinct feature selection sets.

We used 10-fold cross-validation for testing and training. Holdout methods are commonly used when there is a limited quantity of data. In 10-fold cross-validation, the data is divided into 10 approximately equal parts or folds. The first fold is used for testing and folds two through ten are used for training. Each successive fold is used for testing and the remainders for training until all ten iterations are complete. The error rate is calculated for each of the 10 folds and averaged to produce comprehensive error estimation. According to [13], theoretical substantiation and large-scale testing have shown that 10-folds seem to produce prime error estimations.

There is a vast amount of results from our experiment; however, we are only presenting the results of the top three performers. FS15 A6 had slightly better results for precision shown in “Table 5” and correctly classified instances” exhibited in “Figure 2”. FS16 A9 and FS13 A9 had better results for recall, true positives (229) and false negatives (20) and a greater reduction in feature sets: 14 compared to 156 for FS15 A6. FS16 A9 produced better feature selection runtime at 11:37 (min:sec) than FS13 A9 at 17:07 and a significant reduction compared to

FS15 A6 that took over an hour (1:14:25) presented in “Figure 2”. When considering runtime, reduction in features and recall, we have determined that FS16 A9 is the better algorithm to use for this problem. If BV goes undiagnosed and therefore untreated, it can cause very harmful effects 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 highlights the significance of the results as there is an increased chance of pre-term labor and pelvic inflammatory disease (PID).

5. CONCLUSION AND FUTURE WORK

In this paper we conducted experiments using twenty different feature selection algorithms and analyzed the time taken by each of them. We also used nine classification algorithms using the selected features in the previous step and studied the precision and recall of BV disease. Some of the features are not present in many women that lead to minimal effect on overall outcome. On the other hand, the features which were present in all the women have significant effect on the classification results. Our future work will be dedicated towards conducting experiments on the dataset after it has been cleaned, time series removed, clinical data only and medical data only.

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation (NSF): Bio/Computational Evolution in Action Consortium (BEACON) and NIH INBRE award P20GM016454.

REFERENCES

1. Beck D, Foster JA. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PloS one*. 2014; 9:2, e87830.
2. Domingos, P. A few useful things to know about machine learning ACM. New York: 2012. p. 78-87.
3. Hall, MA. Correlation-based feature selection for machine learning. The University of Waikato; 1999.
4. MA B, Forney LJ, Ravel J. Vaginal Microbiome: Rethinking Health and Disease. *Annual Review of Microbiology*. 2012; 66:371–389.
5. Osareh, A.; Shadgar, B. HIBIT 2010 : 5th International Symposium on Health Informatics and Bioinformatics IEEE. Belek, Antalya, Turkey: 2010. Machine learning techniques to diagnose breast cancer.; p. 114-120.
6. Prasad, BDCN.; Prasad, PESNK.; Sagar, Y. A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma). Springer; Berlin Heidelberg, Berlin, Heidelberg: 2011. p. 570-576.
7. Rangari Amit A, Parmjit S, Sharma V. Comparison of the amsel's composite clinical criteria and nugent's criteria for diagnosis of bacterial vaginosis:-a step towards preventing mis-diagnosis. *Journal of Advance Researches in Biological Sciences*. 2013; 5:1, 37–44.
8. 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* 108 Suppl. 2011; 1:11, 4680–4687.
9. Shelton J, Dozier G, Bryant K, Adams J, Popplewell K, Abegaz T, Purrington K, Woodard D, Ricanek K. Genetic based LBP feature extraction and selection for facial recognition. *ACM SE 2011 ACM*. 2011:197–200.

10. 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.
11. Sujatha S, Martin TM, Congzhou L, Frederick AM, Noah GH, Tina LF, Kathy JA, Jeanne MM, David NF. More Than Meets the Eye: Associations of Vaginal Bacteria with Gram Stain Morphotypes Using Molecular Phylogenetic Analysis. *PloS one*. 2013; 8:e78633, 10. [PubMed: 24302980]
12. Williams JA, Weakley A, Cook DJ, Schmitter-Edgecombe M. Machine Learning Techniques for Diagnostic Differentiation of Mild Cognitive Impairment and Dementia. Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence. 2013
13. Witten, IH.; Frank, E.; Hall, MA. Data mining: practical machine learning tools and techniques. third edition.. Morgan Kaufmann Publishers; Burlington, Mass.: 2011.
14. Yau, TMS.; Othman, MFB. Comparison of Different Classification Techniques Using WEKA for Breast Cancer. Springer; Berlin Heidelberg, Berlin, Heidelberg: 2007. p. 520-523.

Figure 1.
Experiment Process

FS15 A6	FS16 A9	FS13 A9
Correctly Classified Instances 97.5016 % Incorrectly Classified Instances 2.4984 %	Correctly Classified Instances 97.3766 % Incorrectly Classified Instances 2.6234 %	Correctly Classified Instances 97.3766 % Incorrectly Classified Instances 2.6234 %
=== Confusion Matrix === a b <-- classified as 227 22 a = yes 18 1334 b = no	=== Confusion Matrix === a b <-- classified as 229 20 a = yes 22 1330 b = no	=== Confusion Matrix === a b <-- classified as 229 20 a = yes 22 1330 b = no
Number of Features: 156 Feature Selection Time: 1:14:25 Classification Time: 0:00:04	Number of Features: 14 Feature Selection Time: 0:11:37 Classification Time: 0:00:02	Number of Features: 14 Feature Selection Time: 0:17:07 Classification Time: 0:00:02

Figure 2.
Results

Table 1

Feature Selection Algorithms

CfsSubsetEval	FilteredSubsetEval
ClassifierSubsetEval	WrapperSubsetEval
ConsistencySubsetEval	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Search Method Algorithms

BestFirst	LinearForwardSelection
RankSearch	SubsetSizeForwardSelection
GeneticSearch	GreedyStepwise

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Classification Algorithms

Code	Algorithm	Code	Algorithm
A1	Bagging	A6	RandomForest
A2	RBFNetwork	A7	LogitBoost
A3	J48	A8	KStar (K*)
A4	NaïveBayes	A9	FT
A5	AdaBoost.M1		

Table 4

Feature Selection Sets

Code	Attribute Evaluator	Search Method	Classifier (For Wrappers)
FS1	CfsSubsetEval	BestFirst	
FS2	CfsSubsetEval	GeneticSearch	
FS3	CfsSubsetEval	RankSearch	
FS4	ClassifierSubsetEval	GeneticSearch	OneR
FS5	ConsistencySubsetEval	BestFirst	
FS6	ConsistencySubsetEval	Genetic Search	
FS7	ConsistencySubsetEval	LinearForwardSelection	
FS8	ConsistencySubsetEval	RankSearch	
FS9	ConsistencySubsetEval	SubsetSizeForwardSelection	
FS10	FilteredSubsetEval	BestFirst	
FS11	FilteredSubsetEval	GeneticSearch	
FS12	FilteredSubsetEval	RankSearch	
FS13	WrapperSubsetEval	BestFirst	NaïveBayes
FS14	WrapperSubsetEval	BestFirst	Bagging
FS15	WrapperSubsetEval	GeneticSearch	Bagging
FS16	WrapperSubsetEval	GreedyStepwise	NaïveBayes
FS17	WrapperSubsetEval	LinearForwardSelection	NaïveBayes
FS18	WrapperSubsetEval	RankSearch	NaïveBayes
FS19	WrapperSubsetEval	SubsetSizeForwardSelection	NaïveBayes
FS20	WrapperSubsetEval	SubsetSizeForwardSelection	Bagging

Table 5

Precision and Recall

	FS15 A6	FS16 A9	FS13 A9
Precision	0.927	0.912	0.912
Recall	0.912	0.92	0.92