Journal of
Advanced Practices in Nursing
ISSN: 2573-0347

Erjaei, Adv Practice Nurs 2017, 2:3
DOI: 10.4172/2573-0347.1000140

Research Article Open Access

Helicobacter pylori Detection Using Machine Learning Algorithm

Erjaei MH*

Department of Computer Science, Wichita State University, USA

***Corresponding author:** Mohammad Hossein Erjaei, Department of Computer Science, Wichita State University, USA, Tel: 3169783456; E-mail: mxerjaei@shockers.wichita.edu

Received date: July 8, 2017; **Accepted date:** August 7, 2017; **Published date:** August 14, 2017

Abstract

It is well known that *Helicobacter pylori (H. Pylori)* is a major cause of chronic active gastritis in both children and adults. There are a variety of tests for detection of *H. pylori* infection, however, in medicine, the only way to diagnose the existence of *H. pylori* microbe is doing endoscopy which is painful and insufferable for young children [1]. To solve this problem, some machine learning classifiers have been used here to diagnose the existence of this infection. As we will see, using machine learning classifier for diagnose the existence of *H. pylori* is an alternative method to avoid painful endoscopy. One hundred patient related data has been used from previous published study. There are twenty features in this dataset, such as: abdominal pain and nausea. We have further investigated the contribution of each single feature by using leave-one-feature-out model, where in each experiment one feature was removed from all features model. This model can help us to see how the features interact and how the most and the least informative features can be found, respectively.

Keywords: *Helicobacter pylori*; Machine learning; Leave one feature out model

Introduction

H. pylori is a type of bacteria which can enter human body and live in digestive tract. After many years, they can cause sores and called ulcers in the lining of human stomach or the upper part of small intestine. For some people, this infection can lead to stomach cancer [1].

Infection with *H. pylori* is common. About two-thirds of the world's population carry it in their bodies. For most people, it doesn't cause ulcers or any other symptoms. Otherwise, if there is a problem, there are some medicines that can kill the germs and help sores heal [2].

Testing for *H. pylori* infection should be considered in patients with a positive family history of gastric cancer, those with refractory iron deficiency anemia, and before long term therapy with proton pump inhibitors [3]. There are varieties of tests for detection of *H. pylori* infection which can be classified as invasive *vs.* non-invasive.

In the invasive test a gastric specimen is obtained through endoscopy and further used for culture, histopathology, PCR, and rapid urease test (RUT). Non-invasive tests include detection of *H. pylori* antigens in stool, detection of antibodies against *H. Pylori* in serum, and urea breath test. The best test for detection of *H. pylori* is one which is available, minimally invasive, greatly accurate and inexpensive [4].

As mentioned before, in medicine, the only way to diagnose the existence of *H. pylori* microbe is doing endoscopy which is insufferable for young children. As an alternative method, Bagherpour et al. [5] used some classifiers to diagnose the existence of this infection. However, it seems that more complex machine learning algorithm can employed to find a better accuracy.

Machine learning is a sub major of computer science and according to Arthur Samuel in 1959, gives "computers an ability to learn without being explicitly programmed" [6]. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible. For example, machine learning can be useful in email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR) and computer vision [7].

Here, we have proposed a non-invasive route by using more complex learning algorithm for estimation of the possibility of *H. pylori* infection, according to the patient's history, gastrointestinal sign and symptoms. Finally, we have used leave-one-feature-out model to further investigate the contribution of each single feature. Indeed, in each experiment one feature was removed from the others to see how this feature effects the result. Doing this, we can find the most and the least informative features, respectively.

Materials and Methods

Data collection

One hundred patient related data was gathered through a randomized clinical trial study, where all children <18 year of age with possibility of *H. pylori* infection; per their sign and symptoms, whom had referred to the Gastrointestinal clinic afflicted to Shiraz University of Medical Sciences from April 2011 till September 2011 were enrolled. First a questionnaire form was completed for each patient, including questions regarding the patient's symptoms (eg. abdominal pain, nausea, vomiting, halitosis, GI bleeding) there duration, positive history of treatment with antacids (H2 blockers and proton pump inhibitors), and any positive family history of acid peptic diseases in their first-degree relatives. Also, all patients were examined for tenderness in their epigastric area and if so this was entered in the form. The patient's weight and height were as well recorded in the questionnaire form. Questions regarding symptoms which could be possibly correlated to *H. pylori* infection in children were derived from previous studies on this concept [5].

Further an endoscopy was performed for all subjects, through which an antral and corpus mucosal biopsy was obtained for histopathology and RUT. Biopsy specimens for histology were fixed in formalin and were sent to Shahid Motahari Pathology Laboratory of Shiraz University of Medical Sciences for analysis. Results regarding the histopathology and RUT were also entered in the form [5].

The used dataset is collected through a six months period from those patients who need to perform the endoscopy in order to diagnose the existence of *H. pylori* infection. The features of dataset are: Male or Female, Abdominal pain, Nocturnal awakening, Nausea, Vomiting, Halitosis, Heart Burn, Bloating, Belching, GI bleeding, Constipation, Diarrhea, Weight loss, Fatigue, Epigastric tenderness, Weight, Height, Duration of symptoms, Previous treatment, Previous Endoscopy, Previous family H Acid peptic Dx, Rapid Urease test before therapy. The Endoscopy feature has been used in our dataset for training algorithms, and later we have tested the algorithms by eliminating this feature from our dataset. This feature has been used just for accuracy measurement and no need of Endoscopy data in real implementation. For more information in data collection please refer to Bagherpour et al. [5]. In what follows, we recall these data as our dataset.

Tools

In this article, Waikato Environment for Knowledge Analysis (WEKA) has used as a tool to apply different machine learning algorithms on the dataset. WEKA is a suite of machine learning software written in Java and developed at the University of Waikato, New Zealand since 1999. It is free software licensed under the GNU Public License. WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these tasks [8].

WEKA supports several standard data mining tasks such as: data pre-processing, clustering, classification, regression, visualization and feature selection. All of WEKA techniques are predicated on the assumption that data is available as one flat file or relation, where each data point is described by a fixed number of attributes. These attributes are ether numeric or nominal, however, some other attribute types are supported too. WEKA provides access to SQL databases using Java database connectivity and can process the result returned by a database query [9]. In order to evaluate the performance of each method on our dataset, we have used cross validation method.

Occasionally, this method called rotation estimation and that is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. Cross validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. The basic form of this validation method is k-fold cross-validation. Other forms of this method are particular cases of k-fold or involve repeated rounds of k-fold cross validation [10].

Support vector machine method

Support Vector Machine (SVM) is the first algorithm which is used in this study. SVM, including support vector classifier and support vector regress or, are among the most robust and accurate methods in all well-known data mining algorithms. SVM, which was originally developed by Vapnik in 1990s, has a theoretical foundation rooted in statistical learning theory, requires only as few as a dozen examples for training, and is often insensitive to the number of dimensions [11].

There are some factors which help SVM to run more accurate. One of the most important factor in SVM is Gamma. Gamma defines how far the influence of a single training example reaches. Low values-far and high values-close (Figure 1).

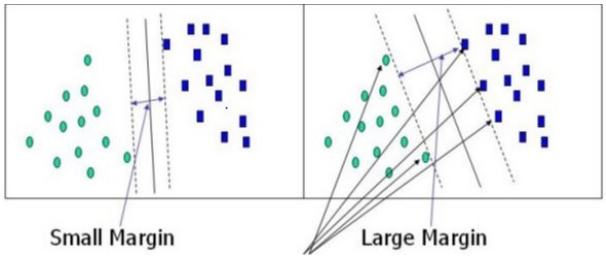

Figure 1: Choosing high value or low value for Gamma in SVM.

In this study to find the best value of Gamma for our dataset, we tried all possible values between 0 and 0.5 by step size 0.01. Table 1 shows various values of Gamma and correct percentages by applying the method on our dataset.

Gamma	Correct	Incorrect
0.01	59.78%	40.22%
0.05	70.65%	29.35%
0.11	73.91%	26.09%
0.13	76.08%	23.92%
0.25	77.17%	22.83%
0.30	73.91%	26.09%
0.35	72.48%	27.52%
0.40	75%	25%
0.50	72.82%	27.18%

Table 1: Some tested gamma value for SVM method.

According to the results shown in Table 1, we are able to provide the best performance of SVM method on our dataset by using Gamma=0.25.

Sequential minimal optimization

Sequential Minimal Optimization (SMO) is an algorithm for solving Quadratic Programming (QP) problem that arises during the training of SVM. It was first invented by John Platt in 1998 at Microsoft Research lab [7]. SMO is widely used for training support vector machines and is implemented by the popular LIBSVM tool [12].

The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers [13]. There are some factors which help SMO to run more accurate.

One of the most important factor in SMO is C. C is similar to Gamma in SVM, however, in order to find the best value for C, we need to test some more range of values. Here, to find the best value of C, we have tried all possible values between 0 and 5 by step size 0.10 on our dataset. Table 2 shows various values of C and their correct percentages, relatively.

C- value	Correct	Incorrect
0.1	69.56%	30.43%
0.3	73.91%	26.08%
0.7	76.08%	23.91%
1.5	77.17%	22.82%
2	78.26%	21.73%
2.5	79.17%	20.83%
3.5	80.43%	19.56%
4	79.34%	20.65%
4.5	78.26%	21.73%
5	77.26%	20.73%

Table 2: Some tested C-value for SMO method.

According to the results shown in Table 2, C=3.5 is a good value to provide the best performance of SMO method on our dataset.

Ensemble

Ensemble learning is a process which multiple models, such as classifiers or experts, are strategically generated and combined to solve a computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one.

Other applications of ensemble learning include assigning a confidence to the decision made by the model can be recall as selecting optimal (or near optimal) features, data fusion, incremental learning, non-stationary learning and error-correcting.

Ensemble based systems can be, perhaps surprisingly, useful when dealing with large volumes of data or lack of adequate data. When the amount of training data is too large and making a single classifier training is difficult, the data can be strategically partitioned into smaller subsets [14].

Using SMO approach as a baseline for classification, we applied two SMO-based ensembles, i.e. Adaptive boosting (adaBoosting) and Bootstrap Aggregating (Bagging) for classifying our data. AdaBoost uses a weighted data sampling and voting scheme. The algorithm starts by building first base classifier, which is trained on the dataset with equal weights.

For the construction of subsequent classifiers, the instances misclassified by the previous classifier are assigned higher weights, while the weights of the instances that are correctly classified remain the same. The weights of all instances in the whole dataset are then normalized so that all weights add up to 1, and then used for sampling for the next classifier.

The final classification for an instance is based on the classifications by all classifiers, with each classifier weighted too. The class with the highest weighted votes is the final classification [15]. Bagging is a method whose classification takes the majority votes of multiple classifiers thus forming a hypothetical "committee".

The training set of each classifier model can be sampled by bootstrap sampling, i.e., randomly selecting a subset of given dataset with replacement, allowing for sample values to be independent of one another [16].

Knowing that two SMO-based ensemble methods have used, we needed to find the best value for C. According to Section 2.4, C=3.5 has used to get the best performance in adaBoosting and Bagging methods.

Leave-one-feature-out model

In machine learning and statistics, feature selection, also known as variable selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

- Simplification of models to make them easier to interpret by researchers/users
- Shorter training times
- To avoid the curse of dimensionality
- Enhanced generalization by reducing over fitting

We further investigated the contribution of each single feature by using leave-one-feature-out model, where in each experiment one feature removed from all feature model. After removing one feature in each experiment, the performance of bagging algorithm has measured by using Receiver Operating Characteristic curve (ROC curve) and F-Measure.

Figures 2-5 show the ROC curves of each model on our dataset. By analysing each feature using the leave-one-feature-out model, we found that abdominal pain feature is the most informative in the dataset.

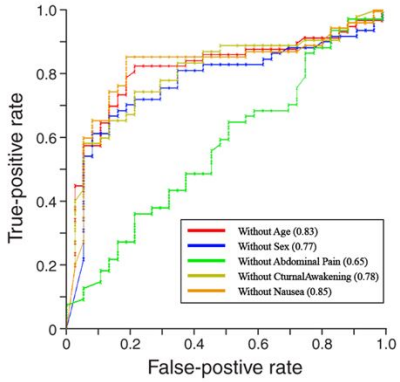

Figure 2: ROC curves for the leave-one-feature-out model using bagging. In this figure, we can see comparison between ROC curves without age, sex, abdominal pain, cturnal awakening and nausea features.

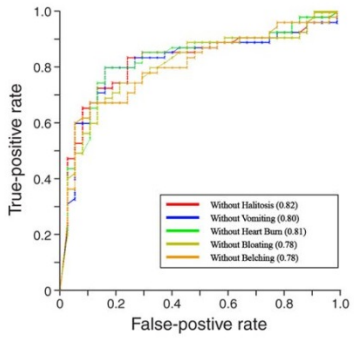

Figure 3: ROC curves for the leave-one-feature-out model using bagging. In this figure, we can see comparison between ROC curves without halitosis, vomiting, heart burn, bloating and belching features.

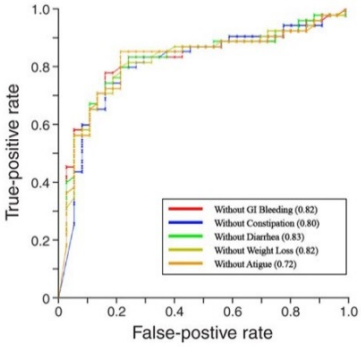

Figure 4: ROC curves for the leave-one-feature-out model using bagging. Here, we can see comparison between ROC curves without GI bleeding, constipation, diarrhea, weight loss and atigue features.

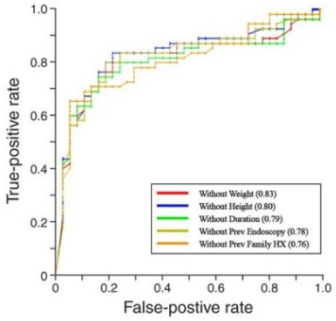

Figure 5: ROC curves for the leave-one-feature-out model using bagging. In this figure, we can see comparison between ROC curves without Weight Loss, height, Duration, Previous Endoscopy and Previous Familu HX.

Results and Discussion

Single learning methods comparison

In this subsection, we have compered the results of applying methods SVM, SMO and some other previously used methods [5] including: Naïve Bayes, Decision Tree, Logistic Regression on our dataset. These results are reported in Table 3.

Method	Correct	Incorrect	F-measure
Naïve Bayes	58%	42%	-
Decision Tree	71.30%	28.70%	-
Logistic Regression	80%	20%	-
SVM	77.17%	22.83%	0.796
SMO	80.43%	19.56%	0.836

Table 3: Performance comparison single learning methods.

As we can see in Table 3, by applying SVM method we will have some improvement with respect to Naïve Bayes. However, comparing with previously studied Logistic Regression method [5], SVM method is weaker. On the other hand, SMO method shows 0.43% improve performance with respect to the logistic regression.

We should pointed out that here we have used WEKA as machine learning tool, while in previous study [5] Statistical Package for the social sciences has used. That is, by using the same pervious study's tool we may achieve the better accuracy.

Ensemble learning methods comparison

In the second set of results, we have illustrated the comparison between applying of SMO method and ensemble methods including adaBoost and Bagging. Using our dataset, these compression results have shown in Table 4.

Method	Correct	Incorrect	F-Measure
SMO	80.43%	19.56%	0.836
AdaBoost	77.17%	22.82%	0.804
Bagging	80.43%	19.56%	0.836

Table 4: Performance comparison between SMO and adaBoost.

According to the results in Table 4, by applying ensemble methods we were not able to improve the performance in comparing with SMO and adaBoost methods. However, no change can be seen in performance by using Bagging method in comparing with SMO method.

Study leave-one-feature-out model

In this study by applying leave-one-feature-out model, we have found that abdominal pain feature is the most informative data in our dataset. By removing this feature from the dataset F-Measure decreased to 0.656 (Figure 6).

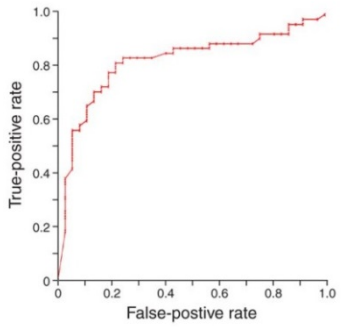

Figure 6: ROC curve of SMO based bagging method for the dataset.

On the other hand, applying this model on our dataset showed that, we could increase the performance of classification by removing the feature nausea. Figure 7 illustrates the comparison of ROC curves with all features in, and those without nausea feature. Here, we should note that Bagging method has been applied to our dataset.

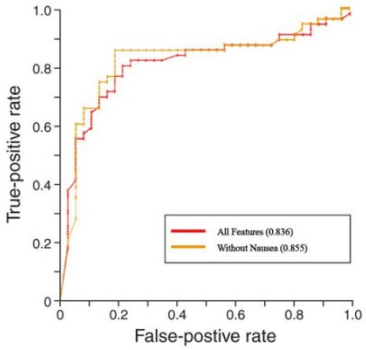

Figure 7: Comparison of ROC curves with all features in, and those without nausea feature.

Our Study shows that by removing nausea feature, we can increase F-Measure up to 0.855 (Table 5).

Method	Correct	Incorrect	F-Measure
Bagging before removing nausea	80.43%	19.56%	0.836
Bagging after removing nausea	82.60%	17.40%	0.855

Table 5: Performance comparison between SMO and SMO based ensemble methods.

Final comparison

Table 6 shows the comparison between the results in pervious study [5] and what we have obtained in this study. As we can see in this Table 6, adaBoost and SVM cannot improve the performance compering with Logistic Regression in previous study. However, SMO and Bagging, by removing nausea feature from dataset, can improve performance up to 0.43% and 2.60%, respectively.

Method	Correct	Incorrect	F-Measure
Naïve Bayes	58%	42%	-
Decision Tree	71.30%	28.70%	-
Logistic Regression	80%	20%	-
adaBoost	77.17%	22.82%	0.804
SVM	77.17%	22.83%	0.796
SMO	80.43%	19.56%	0.836
Bagging (Without nausea)	82.60%	17.40%	0.855

Table 6: Comparison between previous methods and our methods.

Conclusion

In this article, we have presented a comparative study applying several machine learning algorithms for classifying *H. pylori*. Based on two performance evaluation metrics, SMO learning algorithm created reasonable performance. Furthermore, SMO based bagging algorithm were shown to improve the performance over the single SMO algorithm while we used the analyses of leave-one-feature-out model on our dataset. The analyses of the contribution of single features by using leave-one-feature-out model suggests that the most informative feature is abdominal pain in our dataset. Moreover, this analysis shown that removing nausea feature can improve our performance in SMO based bagging algorithm.

In overall, these results suggest that SMO based ensemble learning algorithm can be used to diagnoses *H. pylori* infection. According to our discussion and results in this study, this diagnosis is accurate up to 82.60%. The implementation of this article could be user friendly software. In this case, different users with no background of machine learning and programming can use the software easily with similar performance.

References

1. Drumm B (1993) Helicobacter pylori in the pediatric patient. Gastroenterol Clin North Am 22: 169-182.
2. Holcombe C, Omotara BA, Eldridge J, Jones DM (1992) H. pylori, the most common bacterial infection in Africa: A random serological study. Am J Gastroenterol 87: 28-30.
3. Bourke B, Ceponis P, Chiba N, Czinn S, Ferraro R, et al. (2005) Canadian helicobacter study group consensus conference: Update on the approach to Helicobacter pylori infection in children and adolescents-An evidence-based evaluation. Can J Gastroenterol 19: 399-408.
4. Guarner J, Kalach N, Elitsur Y, Koletzko S (2010) Helicobacter pylori diagnostic tests in children: Review of the literature from 1999 to 2009. Eur J Pediatr 169:15-25.
5. Bagherpour M, Erjaee A, Rasekh AH, Dehghani SM (2014) Data mining applications in a medical system: A Case Study. Encyclopedia of Business Analytics and Optimization: 602-608.
6. Munoz A (2014) Machine Learning and Optimization.
7. Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC (2010) Machine learning in medical imaging. IEEE signal process mag 27: 25-38.
8. Witten IH, Frank E (2006) Data Mining: Practical machine learning tools and techniques. Biomed Eng 5: 51.

9. Reutemann P, Pfahringer B, Frank E (2004) A toolbox for learning from relational data with propositional and multi-instance learners. Adv Art Intel 1017-1023.
10. Refaeilzadeh P, Tang L, Liu H (2009) Cross validation. Encyclopedia of database systems: 532-538.
11. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, et al. (2008) Top 10 algorithms in data mining. Knowledge and information systems 14: 1-37.
12. Platt J (1998) Sequential minimal optimization: A fast algorithm for training support vector machines. 98: 21.
13. Rifkin RM (2002) Everything old is new again: A fresh look at historical approaches in machine learning. Massachusetts Institute of Technology.
14. Baruque B, Corchado E (2011) Fusion methods for unsupervised learning ensembles. Springer: 322.
15. Freund Y, Schapire RE (1995) A desiciontheoretic generalization of on-line learning and an application to boosting. European conference on computational learning theory. pp: 23-37.
16. Breiman L (1996) Bagging predictors. Machine learning 24: 123-140.