IET Software

Special Issue: Advances in Knowledge and Information Software Management

Prediction of leptospirosis cases using classification algorithms

ISSN 1751-8806
Received on 31st July 2016
Revised 9th December 2016
Accepted on 29th December 2016
E-First on 31st January 2017
doi: 10.1049/iet-sen.2016.0193
www.ietdl.org

Nivison Ruy Rocha Nery Jr[1,2], *Daniela Barreiro Claro*[2] ✉, *Janet C. Lindow*[3]

[1]*Gonçalo Moniz Institute (IGM) Oswaldo Cruz Foundation, Ministry of Health, Salvador, Brazil*

[2]*Semantic Formalisms and Applications Research Group (FORMAS), Computer Science Department, Federal University of Bahia (UFBA), Salvador, Brazil*

[3]*Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA*

✉ *E-mail: dclaro@ufba.br*

Abstract: Leptospirosis is a potentially life-threatening disease primarily affecting low-income populations, with an estimated annual incidence of 1.03 million infections worldwide. This disease has symptoms often confused with other febrile syndromes, such as dengue fever, influenza and viral hepatitis, often making diagnosis challenging. Improving the accuracy of early diagnosis of patients with leptospirosis will increase the speed of appropriate antibiotic treatment delivery, and both will improve clinical outcomes for this potentially fatal disease. The authors conducted an analysis of clinically and epidemiologically defined leptospirosis cases to predict disease using data mining classification algorithms. They conducted four sets of experiments to evaluate the performance of the algorithms, assessing their predictive accuracy of using different training and test datasets. The JRIP algorithm achieved 84% sensitivity using a dataset of only confirmed leptospirosis cases, and a specificity of 99% using a dataset of only confirmed dengue cases. Therefore, the approach successfully predicted leptospirosis cases, differentiated them from similar febrile illnesses, and may represent a new tool to assist health professionals, particularly in endemic areas for leptospirosis, accelerating targeted treatment and minimising disease exacerbation and mortality.

1 Introduction

Leptospirosis, a neglected tropical disease, is an acute febrile disease that affects populations living in tropical regions of the world, which can cause potentially fatal pulmonary haemorrhage or acute renal failure. The causative agent of this disease, a bacterium called *Leptospira* [1], is transmitted by direct contact with sewage, water or soil contaminated with urine from infected animals, particularly sewer rats in urban slum settings.

The clinical presentation of leptospirosis patients often includes symptoms that are common to other febrile syndromes such as influenza, dengue fever and viral hepatitis, which hinders accurate diagnosis by health professionals. Often when there is suspicion of leptospirosis, many regions lack confirmatory laboratory testing, and even when possible, the gold standard diagnostic tests generally take several days, skilled personnel and paired blood samples for accurate diagnosis. This slows notification to public health services [2] and medical personnel, potentially delaying the start of appropriate antibiotic treatment, which is associated with poorer clinical outcomes.

In Brazil and many other regions, the laboratory tests most commonly used to diagnose leptospirosis are the enzyme-linked immunosorbent assay-immunoglobulin M (ELISA-IgM) and micro-agglutination test (MAT) [3]. A major drawback of these assays is that they often require paired blood samples for the highest sensitivity. For instance, the early acute blood sample test results often are negative, and one cannot rule out the diagnosis of leptospirosis [3]. Even with paired blood samples, some cases are only confirmed by quantitative polymerase chain reaction, which detects bacterial nucleotides in whole blood, or by hemoculturing methods (detects the bacteria) [3]. Owing to these significant limitations in the current diagnostic tests, the use of a prediction model based on epidemiological and clinical data may provide a higher predictive value to accurately and rapidly diagnose leptospirosis and aid in the effective treatment of individuals, decreasing the likelihood of the development of more severe disease.

The Gonçalo Moniz Institute (IGM/FIOCRUZ-BA) is an institution linked to the Brazilian Ministry of Health that performs studies on various infectious diseases. Since 1996, the IGM/FIOCRUZ-BA has had a research group dedicated to performing laboratory, clinical and epidemiological studies of leptospirosis, in Salvador, Brazil. These studies include active hospital surveillance at Hospital Couto Maia (HCM), which specialises in infectious diseases; ambulatory monitoring for acute leptospirosis at the Pau da Lima neighbourhood São Marcos health clinic; and epidemiological studies of the natural history of leptospirosis in a Pau da Lima/São Marcos community cohort of ∼14,000 individuals.

The aim of the current study was to analyse whether classification models applied to clinical and epidemiological patient data could accurately identify cases of leptospirosis. This work provided two main contributions:

- Evaluating prediction algorithms to better fit epidemiology and clinical parameters.
- Improving disease identification for health professionals, thereby accelerating the initiation of appropriate treatment of patients.

This work is structured as follows: Section 2 presents related works. Section 3 introduces the methodology with classification models applied to leptospirosis, information regarding the clinical and epidemiological database and techniques for the preprocessing of data, description of the analysed algorithms and evaluation techniques. Section 4 explains the presentation of experiments and results detailing the metrics analysed and the performance of each algorithm. Section 5 describes the experiments. Section 6 discusses the conclusions and future works.

2 Related work

There are various works in the literature related to the knowledge discovery in databases (KDD) applied to health data [4–11]. To perform knowledge discovery, KDD has various steps, as shown in

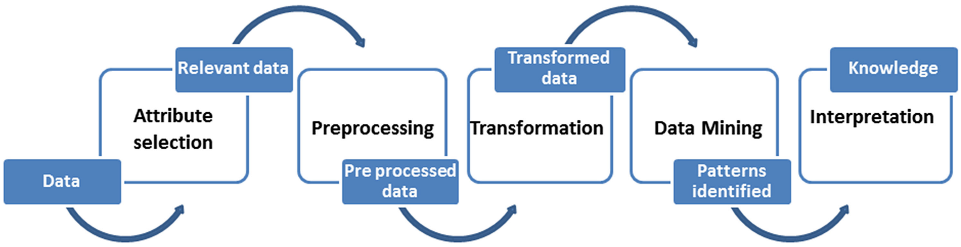

Fig. 1 *KDD PS adapted from Fayyad et al. [8]*

Fig. 1 [8]. The steps attributes selection, processing and transformation are essential for the preparation of data for the step of data mining (DM) [12], particularly when working with incomplete clinical information such as hospitalised patient data [4].

The analysis theme of prediction models has been used by many researchers to identify a variety of diseases using DM. All cited works contained analysed techniques, algorithms and the import of the research. For example, in Garcia *et al.* [4], the authors compared prediction algorithms k-nearest neighbours (KNN), decision tree, logistic regression and support vector machine (SVM) with data imputation methods to identify the best survival prediction model of cases of breast cancer over 5 years of historic data. The best prediction model was obtained with the application of the KNN algorithm with 81% accuracy and 0.78 area under the receiver operating characteristic (ROC) curve.

Yeh *et al.* [5] applied DM techniques to biochemical data from dialysis patients to predict the likelihood of hospitalisation and found that immediate treatment associated with a lower hospitalisation rates. Using decision tree and association rules algorithms, they further determined that the albumin index was an important factor in predicting hospitalisation of patients with an accuracy ranging from 71 to 100% [5].

Fayyed *et al.* [7] constructed a predictive model to improve the diagnosis and prognosis of cardiovascular disease. The authors analysed the algorithms decision tree, Bayesian network and neural networks with the cross-validation (CV) playing technique with 10 folds. The algorithm that showed the best performance for the cases of disease classification was the decision tree with accuracy and sensitivity of >99% [7].

Like leptospirosis, which kills millions of people around the world, malaria is another febrile disease with high mortality rates, especially in Africa. The main reasons are the shortage of health professionals and hospitals equipped for appropriate detection and treatment, especially in rural areas. Oguntimilehin [13] carried out a review of the work in the area of development of predictive models for assistance in the diagnosis and medical treatment of malaria. Fifteen studies were presented with different methodologies supporting early disease diagnosis, which reduces mortality. The author emphasised the importance of having computational tools that could aid in disease diagnosis, but warned of the need to evaluate the accuracy and reliability of these tools before putting them into clinical use [13]. Of particular relevance to the work presented here, Ugwu *et al.* [14] used decision tree as part of the diagnosis and treatment recommendation for malaria. This is a specific example of a well-designed algorithm serving as an important health tool in a region with a shortage of specialised professionals.

In the work of Sahle [9], the reduction of malaria mortality rates was also the motivation for the use of DM techniques. In this work, J48, JRip and MLP algorithms were used for the methods decision tree, classification rules and neural networks, respectively. The experiments showed promising results for models evaluated with an accuracy rate of 97% cases for MLP and 96% for the models obtained with J48 and JRip [9]. Thus, this strategy represents a promising methodology for early disease diagnosis, particularly in global regions lacking in adequate medical support.

Bakar *et al.* [10] developed a dengue outbreak detection model using multiple classifiers based on rules, such as decision tree, and naive Bayes classifier. They found the use of multiple classifiers improved the accuracy and quality of rules in comparison with a single classifier. For example, the best results for single classifier analysis resulted in an ROC curve of 0.729, while experiments using multiple classifiers yielded an ROC curve of 0.761.

Currently, only one study applied DM to leptospirosis, though this study was performed in dogs, not in humans [11]. The authors applied the decision tree algorithm to epidemiological and serological data to assess the risk factors for canine leptospirosis and found that while the specificity of J48 algorithm was 87%, the sensitivity of the algorithm was only 66%. Unlike human leptospirosis, where the diagnosis is hampered by symptoms common to other febrile syndromes, leptospirosis is suspected in dogs more frequently because of greater exposure to bacterial reservoirs. To the best of our knowledge, our study is the first to apply these methods to human cohorts with and without laboratory-confirmed leptospirosis and incorporating epidemiology data.

3 Methodology

The KDD process was applied to better classify cases of leptospirosis from the IGM/FIOCRUZ-BA patient and community cohort leptospirosis database. We first determined the type of data available for the disease and defined the limitations of the information database before selecting attributes to extract relevant data (step 2). In the third step, we performed the data preprocessing: cleaning and standardising the existing data. In the fourth stage, we transformed the data into a readable format. We then applied the DM classification algorithms based on three methods (classification rules, a decision tree and Bayesian classification) to predict the learning historical data and cases of the disease. Finally, we compared the results from the algorithms applied and identified which best predicted the statistical parameters available in the classification tool [15].

3.1 Prediction model development

This work was divided into two objectives:

i. *Evaluating the leptospirosis case prediction algorithms of the main classification methods*: Bayesian, classification rules and decision tree. Assessing the validation techniques of percentage sprit (PS) and CV.
ii. *To determine the sensitivity and specificity of the best algorithms identified in independent datasets*: an independent set of patients identified during active hospital surveillance, a dataset containing only leptospirosis confirmed cases, and a dataset of dengue confirmed cases.

In this work, we built on our prior findings in which classification models were compared to determine the best predicted cases of leptospirosis using clinical and epidemiological data [16]. Clinical data consisted of information gathered from hospital records, which included patient history, clinical disease course, performed examinations and treatments. Epidemiological data included information about daily activities of the individual, whether at home or at work to: (i) assist in clinical diagnosis of suspected cases, since many diseases have similar initial symptoms, (ii) identify patterns of disease occurrence and the factors influencing them and (iii) contribute to the case notification system for health departments to facilitate vaccination and disease control campaigns. In the case of leptospirosis, information related to contact with sewage, mud, flooding, garbage or rats in the 30 days

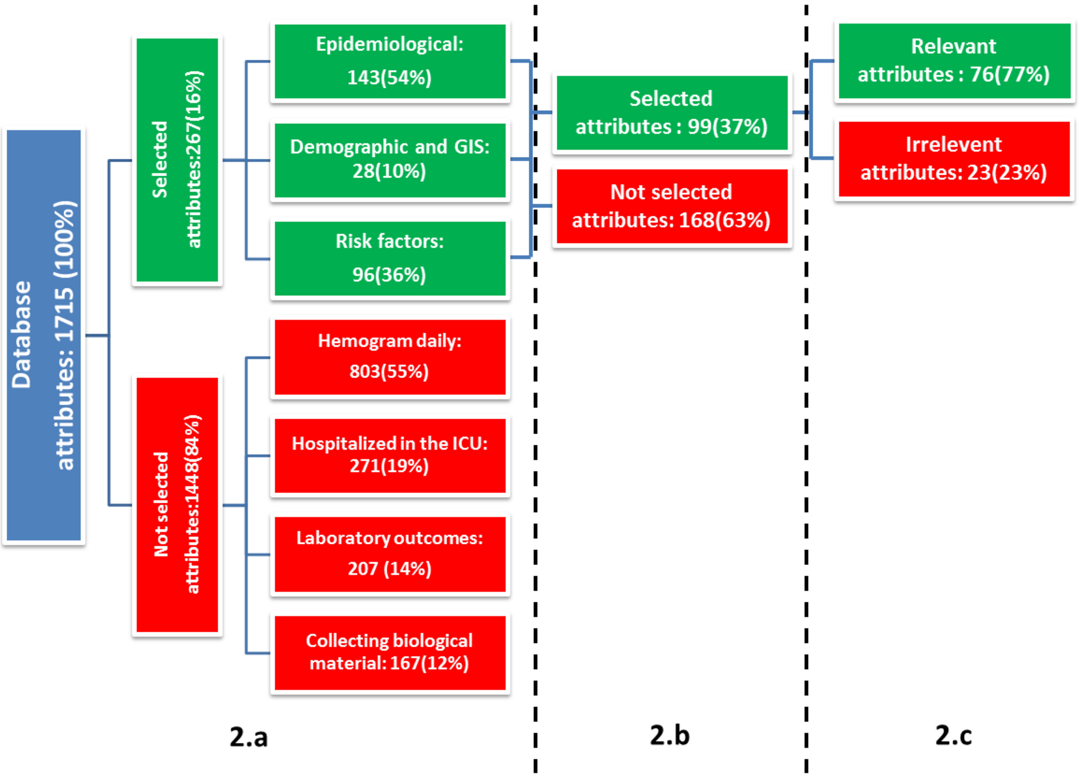
Database attributes: 1715 (100%)
Selected attributes:267(16%)
Not selected attributes:1448(84%)
Epidemiological: 143(54%)
Demographic and GIS: 28(10%)
Risk factors: 96(36%)
Hemogram daily: 803(55%)
Hospitalized in the ICU: 271(19%)
Laboratory outcomes: 207 (14%)
Collecting biological material: 167(12%)
Selected attributes : 99(37%)
Not selected attributes: 168(63%)
Relevant attributes : 76(77%)
Irrelevent attributes: 23(23%)
2.a
2.b
2.c

Fig. 2 *Flow of feature selection step (2.a), pre-processing (2.b) and relevance of the assessment by expert (2.c)*

prior to onset of symptoms was critical in the prediction of clinical suspicion of the disease.

3.2 Leptospirosis database

All patients and community cohort volunteers provided written, informed consent in their native language to participate in the studies described here. Since 1996, the IGM/FIOCRUZ-BA Leptospirosis group has performed active surveillance to identify suspected leptospirosis cases at the HCM in Salvador, Brazil [17, 18]. This study used data collected during hospital surveillance, composed of information collected from patient interviews, medical charts review and laboratory test results. Data from hospital surveillance consisted of

- Demographic data, including home location, age and sex of the participant.
- Epidemiological data collected at the time of patient hospital admission to determine whether patients had risk factors for leptospirosis, and initial results of non-specific examinations, such as complete blood count, chest X-ray and urine output summary. Following hospital discharge, the team collected epidemiological data on final clinical outcome and relevant medical chart history.
- Data related to the past 30 days of activities to identify individual risk factors such as possible exposures (contact with mud, sewage and/or waste), socioeconomic characteristics and/or living or working in a hazardous area.
- Data on household risk factors from the participants' homes: information such as proximity of sewage, garbage accumulation, number of rats seen in/near home, vegetation and animals.
- Georeference data on the exact location of the participant's home for later spatial analyses.

3.2.1 Data preprocessing: The hospital surveillance database consisted of 1715 attributes, with 4675 instances collected from March 1996 to September 2014. The major limitation of this database was that most of the attributes were created in the last 5 years; thus, many variables were missing over 13 years of data resulting in incomplete data.

Since the objective of this model was to predict the diagnosis of leptospirosis based on clinical and epidemiological data without corresponding laboratory data, some attributes such as laboratory results which confirm the diagnosis were removed. Additionally, attributes for <2 years of input data, i.e. biological material collections, clinical and laboratory outcomes, daily blood counts and ICU results were also removed. Thus 267 attributes remained were initially evaluated, as shown in Fig. 2.a.

Then we analyse the missing values for each attribute, removing each variable that outperforms 80% of missing data. We remain with 99 attributes, as depicted in Fig. 2.b. Our resulting dataset was organised on 84% nominal and 16% discrete attributes.

A healthcare professional performed a manual revision of the 99 attributes to remove irrelevant or clinical diagnostic data. This step removed variables lacking clinical relevance, such as whether a particular sample was collected. In addition, clinical diagnostic data were removed to prevent biasing the results of our analysis. As a result of this review, our database contained 76 clinical and epidemiological variables including 80% nominal and 20% distinct attributes, as shown in Fig. 2.c.

Our IGM/FIOCRUZ-BA database consisted of 4675 instances of suspected leptospirosis cases: 2046 were laboratory confirmed (44%) and 2629 were unconfirmed (56%). All the 99 attributes from the initial set with 4675 instances was divided into training and testing data. From the 2629 unconfirmed leptospirosis cases, 596 instances were laboratory confirmed for Dengue. These 596 were used to evaluate the specificity of JRIP and J48 algorithms. The laboratory confirmation was used to validate the effectiveness of our models.

Finally, we created a predict attribute based on laboratory diagnostic tests called *LEPTO* whose response options were *confirmed* or *not confirmed*.

3.3 Analysed algorithms

Our objective for mining the database of hospital surveillance of leptospirosis (IGM/FIOCRUZ-BA) was to predict the disease cases and their relationship to laboratory diagnosis. In the predictive activity, we selected algorithms from the main methods of classification task: decision tree with the algorithms J48 and REPTree; classification rules, with JRip, OneR, PART and DecisionTable (DT) algorithms; Bayesian classification, with the

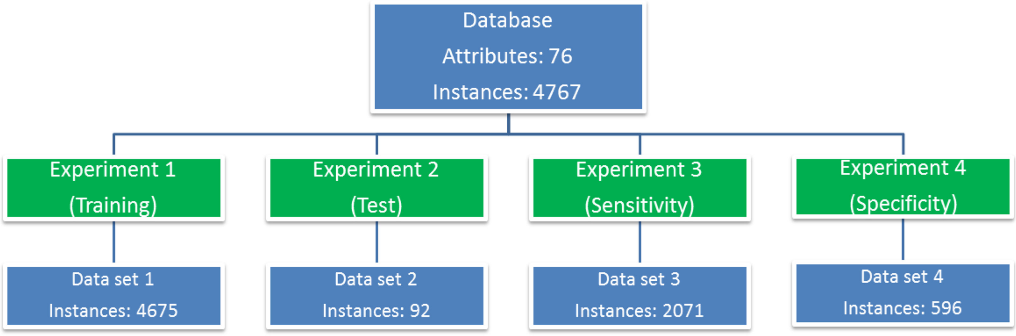

Fig. 3 *Flow experiments and datasets used*

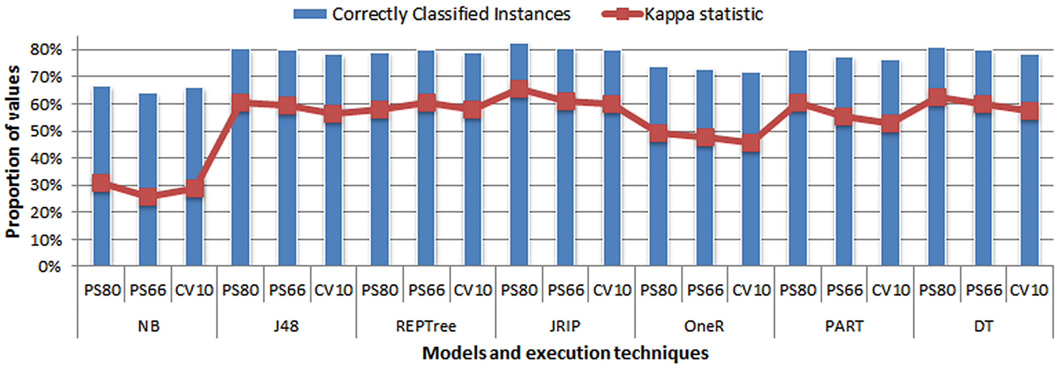

Fig. 4 *Comparison of accuracy of instances and the value obtained with the kappa statistic, by algorithms and implementation techniques [16]*

naive Bayes algorithm (NB) [19]. We choose these methods because they performed well in related work and provided simple, flexible and easily interpretable results.

3.4 Evaluation metrics

To identify the best classification model for leptospirosis, we used statistical techniques to evaluate the metrics. Initially, the confusion matrix is used as the basis for the statistical evaluation parameters of classification models, measuring the number of correct classifications based on disease outcome class, compared to predicted classifications for the algorithms. Of the available statistical metrics, we analysed the sensitivity and specificity of the database instances and kappa statistic, a measure that verifies the relevance of each attribute. The greater the displayed kappa value, the greater the confidence in the classification result. We also evaluated the accuracy of the metrics of classification available using other methods: true positives(TP); true negatives (TN); false positives (FP); false negatives (FN); F-measure, a measure that verifies the performance combining positive predictive value (PPV, PPV = TP/[TP + FP]) and sensitivity values of a rule; Matthews correlation coefficient (MCC), a balanced measure that can be used even when there are disparities in the size of study classes; and ROC area, metric that represents the variation between the true and false-positive rates [20].

Sensitivity is a measure used to evaluate the effectiveness of an instrument to confirm individual with the disease among patients (Sensitivity = TP/[TP + FN]) [21].

Specificity is an assessment that measures how often a method successfully identifies an individual who does not have the disease [21], in this case, leptospirosis (specificity = TN/[TN + FP]).

4 Experiments and results

In this work, we performed four experiments, using specific datasets created for training, testing, evaluating sensitivities on a dataset within confirmed cases and other within only unconfirmed cases, as can be seen in Fig. 3. We describe each experiment followed by each result.

4.1 First experiment

This first experiment aimed to predict cases using our leptospirosis database, which was divided into training and testing datasets. We ran each algorithm three times with different fragmentation techniques: on the first run, we used the technique PS, dividing the dataset into 66% for training and the 34% remaining for testing. On the second run, we applied the PS technique but used 80% for training and 20% for the test. In the third run, we used the technique CV with 10 folds [16].

Tables 1 and 2 [16] show the results separated by algorithms and methodology. Table 1 contains the analysis of the models obtained from NBs, J48 and REPTree, while Table 2 presents the analysis of the models obtained with the JRIP algorithms, OneR, PART and DT. As shown in the tables, we ran each algorithm three times with different techniques. Fig. 4 shows the proportion of correct classifications and their respective values obtained through kappa statistics. While the DT models and J48 yielded promising results, the JRIP algorithm model yielded the best results for accuracy, especially in the percentage split 80 experimental technique.

Tables 1 and 2 show the results of the metrics we considered during the evaluation of the models. Based on the kappa statistics, the model that showed the highest confidence index in all methods of experimentation was the JRip: 0.657 for technical PS 80, 0.612 for PS66 and 0.599 for CV 10. The worst performing model was the NB, with values: 0.310 for PS80, 0.257 for PS66 and 0.289 for CV 10 [16].

Analysing the accuracy of TP models, the model that stood out was the OneR, with 0.946 hit ratio for PS80, 0.938 for PS66 and 0.937 for CV 10. However, among the FP ratings, OneR yielded the highest false-positive rates: 0.432 for PS80, 0.442 for PS66 and 0.451 for CV 10. Thus, this classifier showed high sensitivity for patients with confirmed disease and had the highest test specificity. Table 3 [16] shows the confusion matrix obtained by running the algorithm OneR through technical PS80 with 95% sensitivity, identifying 401 of the total 424 individuals with the disease, but also identifying 290 total of 511 of the group of individuals without the disease (57% specificity), damaging the confidence model.

Different from the OneR algorithm results, the model obtained by JRIP yielded a balanced sensitivity and specificity of 85 and 81%, respectively, a reduction in sensitivity, but a notable increase

Table 1 Analysis of statistics of Bayesian algorithms and decision trees, comparing the PS experimental models (Split 66 and 80) and CV 10 folds [16]

Metrics	Bayesian			Decision tree					
	NB			J48			REPTree		
	PS80	PS66	CV10	PS80	PS66	CV10	PS80	PS66	CV10
accuracy	0.667	0.643	0.662	0.804	0.799	0.785	0.791	0.803	0.79
kappa statistic	0.31	0.257	0.289	0.605	0.596	0.563	0.582	0.605	0.579
TP rate	0.474	0.423	0.466	0.788	0.794	0.759	0.802	0.811	0.808
FP rate	0.172	0.174	0.186	0.182	0.196	0.195	0.217	0.203	0.224
F-measure	0.696	0.67	0.66	0.782	0.772	0.75	0.754	0.769	0.737
MCC	0.325	0.274	0.3	0.605	0.596	0.563	0.582	0.605	0.581
ROC area	0.774	0.756	0.77	0.848	0.828	0.811	0.847	0.847	0.848

Table 2 Analysis of the statistical measures results of classification rules algorithms, comparing the PS experimental models (Split 66 and 80) and CV 10 folds [16]

Metrics	Classification rules											
	JRIP			OneR			PART			DT		
	PS80	PS66	CV10	PS80	PS66	CV10	PS80	PS66	CV10	PS80	PS66	CV10
accuracy	0.829	0.806	0.801	0.739	0.731	0.718	0.803	0.778	0.767	0.812	0.801	0.787
kappa statistic	0.657	0.612	0.599	0.494	0.477	0.459	0.605	0.552	0.529	0.625	0.6	0.572
TP rate	0.851	0.842	0.815	0.946	0.938	0.937	0.816	0.734	0.748	0.863	0.79	0.806
FP rate	0.19	0.225	0.21	0.432	0.442	0.451	0.207	0.185	0.217	0.231	0.189	0.228
F-measure	0.788	0.757	0.75	0.645	0.639	0.616	0.765	0.768	0.727	0.756	0.777	0.733
MCC	0.659	0.615	0.6	0.541	0.524	0.509	0.606	0.552	0.529	0.63	0.6	0.574
ROC area	0.844	0.827	0.826	0.757	0.748	0.743	0.829	0.82	0.798	0.863	0.83	0.843

Table 3 Matrix confusion OneR model with experimental technique PS80 [16]

Class	No disease	With disease
no disease	290 (57%)	221 (43%)
with disease	23 (5%)	401 (95%)

Table 4 Matrix confusion JRip model with experimental technique PS80 [16]

Class	No disease	With disease
no disease	414 (81%)	97 (19%)
with disease	63 (15%)	361 (85%)

in specificity. Table 4 [16] shows the confusion matrix obtained by running the algorithm JRIP through PS80 technique, indicating that the model identified 361/424 individuals with the disease and was also effective in identifying individuals without the disease, 414/511.

Evaluating the F-measure, the JRip model showed better results in the PS80 and CV 10, with values 0.788 and 0.750, respectively, but the DT was the best performer (0.777), followed by J48 (0.772) and the PART (0.768) for the PS66 model. The JRip was the fifth best using the PS66 method. Based on F-measure, the JRIP had a sensitivity of 84% and specificity of 75%, while the J48 had a lower sensitivity (75%) and similar specificity (79%).

Analysing the MCC, we observed that the JRip model outperformed all other models with values of 0.659 for PS80, 0.615 for PS66 and 0.600 for CV 10.

Finally, we observed that PS66 method was slightly better compared to the CV 10, but it reduced the accuracy on PS80.

Among the models analysed, JRip presents the best performance in applying metrics especially in the proportion of correct classification of instances, kappa statistics and F-measure for almost all testing methods, and best performances in PS80. JRip produced correct classifications rate of ~80%. The Bayesian model performed worst for all methods of experimentation with the best kappa value of 0.310 and F-measure as of 0.696. Additionally, we analysed the performance of the DM methods for classifying the disease. Among the implemented methods, classification rules were the most efficient, mainly due to the performance of JRip models and PART. The Bayesian method showed the lowest performance in the analysis [16].

4.2 Second experiment

The JRIP and J48 algorithms yielded the best results in the first experiment and were therefore selected for additional experiments.

To further test our model, we added patients' records from October 2014 to June 2016 to the database:

i. *Training set*: the same dataset used in the first experiment. This dataset initially had 99 variables; however, a manual review of all 99 attributes was conducted with a healthcare professional to remove irrelevant information. At the end of this step, there were 76 variables remaining and 4675 instances of recruited patients from March 1996 to September 2014.
ii. *Test set*: containing 76 variables and 92 instances of patients recruited from October 2014 to June 2016.

In test experiments, the model obtained with the JRIP algorithm retained the sensitivity of 84%, but the specificity decreased to 52%. The specificity of the model obtained with J48 in this dataset was also 52% though the sensitivity increased from 75 to 82% compared to Experiment 1.

4.3 Third experiment

The third experiment aimed to evaluate exclusively the sensitivity of our models obtained with JRIP and J48 algorithms in the training set. We tested two models on a dataset containing only clinically confirmed leptospirosis cases. This dataset contains 76 variables and 2071 instances of patients recruited from March 1996 to June 2016.

In this experiment, the sensitivities between the JRIP model and the J48 model on the training set were compared: JRIP minimally reduced the specificity from 84 to 82% while J48 increased the specificity from 75 to 85%.

4.4 Fourth experiment

The aim of this experiment was to assess the specificity of JRIP and J48 algorithms using a database containing only clinically

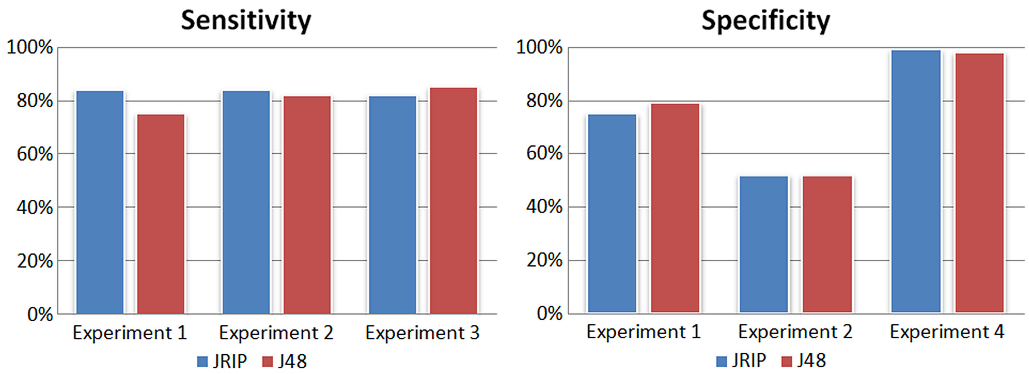

Fig. 5 *Comparison of experiments for sensitivity and specificity*

confirmed dengue cases, a disease with initial symptoms similar to leptospirosis. This dataset contained 76 variables and 596 instances of patients recruited from March 1996 to June 2016.

Both experiments displayed relevant results. The JRIP and J48 experiments had specificities of 99 and 98%, respectively, indicating that the models have high specificity in discriminating among patients with and without leptospirosis.

Fig. 5 shows a comparison between sensitivity and specificity rates. This dataset only contains dengue cases, we do not evaluate the sensitivity for this experiment. Likewise, Experiment 3 does not report specificity data, since the dataset only contains leptospirosis cases.

5 Discussion

Leptospirosis is a neglected disease of global import that occurs primarily in rural or urban areas lacking quality sanitation systems [22]. Rapid diagnosis and appropriate treatment of leptospirosis significantly improve the outcome of this potentially fatal disease. Since the clinical presentation of leptospirosis is often difficult to differentiate from other co-circulating febrile diseases, diagnosis of leptospirosis can be challenging. The use of a classification model that accurately predicts whether an individual with suspected leptospirosis should be treated for leptospirosis would enhance diagnosis and appropriate medical treatment of cases lacking laboratory data and complete clinical data. Additionally, the model can be applied to databases containing historical data facilitating the estimation of an average number of future cases, thereby, informing intervention and control strategies for leptospirosis in resource-limited regions.

Using clinical and epidemiological data from long-term active hospital and ambulatory surveillance for leptospirosis cases in Brazil, we ascertained the JRip algorithm is the best prediction model for identifying leptospirosis cases, with 84% accurate classification of disease cases, defined by laboratory confirmation, followed by J48 (82%). Our results provide strong evidence that individuals with leptospirosis can be identified using the prediction model based on limited clinical and epidemiological data, in regions with other febrile illnesses presenting with similar symptomology.

The data used as reference values to assess the correct classification of the analysed models were obtained through specific laboratory tests for predicting leptospirosis. According to the World Health Organization (WHO) [3], laboratory confirmation of the disease should preferably include two blood samples collected at different times: the first during the acute phase of the disease and the second 14–21 days following the first blood draw for ELISA and MAT testing. The paired samples are needed because many acute phase samples are negative for these antibody-dependent assays complicating disease diagnosis. This prolongs laboratory confirmation of the diagnosis. Thus, our methodology, combined with current practices, could help improve accurate diagnosis and greatly speed the application of appropriate antibiotic treatment.

Of the 2071 confirmed leptospirosis cases in the IGM/FIOCRUZ-BA database, 13% (271/2071) were negative by ELISA [3] and 27% (560/2071) were negative by MAT [3] when only the early acute blood sample was analysed. Unlike the findings for the predictive model for canine leptospirosis [11], we found the JRip algorithm yielded the highest sensitivity (85% 230 of the 271 confirmed by ELISA and 476 of the 560 confirmed by MAT) for predicting human leptospirosis. It represents a potential new diagnostic tool for diagnosis disease, particularly in resource-limited settings.

6 Conclusion

In this paper, we employed DM techniques, specifically classification rules, decision trees and Bayesian network algorithms to identify the best model for the prediction of leptospirosis in patients. In our experiments, classification rules algorithms presented high accuracy compared with other tasks analysed. Additionally, we evaluated data fragmentation techniques: PS and CV, whose purpose is to divide the database into training and testing data [16].

In our analyses of the classification models, the JRip rating rules method produced better results in the evaluated metric. The model obtained with JRip showed better accuracy compared with other models or the fragmentation techniques PS and CV. The PS experimental model presented the best performances, in particular with the increased proportion of instances for training, PS 80% [16]. Despite considerable success using the validation technique PS, it is not the most commonly used since the proportion used for training and testing can be biased by various situations, i.e. a proportion of recruited participants with a given outcome can be the basis for training or testing. The most suitable technique is the CV with k folds. In this work, we used $k = 10$, where the data was divided into training and testing, and then evaluated the accuracy of the model, repeating the iteration ten times and the model accuracy was averaged to represent the data [23].

The models obtained with JRIP and J48 algorithms identified suspected leptospirosis cases with reasonable sensitivities of 84 and 82%, and high specificities of 99 and 98%, respectively. The models identified suspected cases of leptospirosis and presented great rates in ruling out of other febrile diseases often confused with leptospirosis, such as dengue fever, used as control patients for this work.

The incorporation of the JRip or J48 models by health care professionals in hospitals, research centres and in epidemiological surveillance services to predict cases of leptospirosis is crucial as a new diagnosis aid for clinician decision-making. The earlier the accurate diagnosis of leptospirosis is made, the faster the individual will be treated, reducing the likelihood of progression to severe disease outcomes such as pulmonary haemorrhage or death.

In future, we intend to expand the comparisons of algorithms by evaluating other prediction methods, such as KNN, SVM and logistic regression. We also will assess the prediction model of leptospirosis obtained by DM and validate the results with inferences made in an ontology for the disease that is currently under development, verifying the efficiency and effectiveness compared with semantic relations.

7 Acknowledgment

The authors thank the Yale-Fiocruz Leptospirosis group and the Laboratory of Pathology and Molecular Biology of the Gonçalo Moniz Institute-IGM/FIOCRUZ-BA for the availability of data and information on the disease, and important specialists, Dr. Albert Ko, Dr. Mitermayer Galvão dos Reis and Dr. Federico Costa for helpful discussions.

8 References

[1] Reis, R.B., Ribeiro, G.S., Felzemburgh, R.D*., et al.*: 'Impact of environment and social gradient on Leptospira infection in urban slums', *PLoS Negl. Trop. Dis.*, 2008, **2**, (4), p. e228

[2] Brasil. Ministerio da Saude: 'Guia de vigilancia epidemiologica / Ministerio da Saude, Secretaria de Vigilancia em Saude, Departamento de Vigilancia Epidemiologica' (Ministerio da Saude, Brasilia, 2009, 7th edn.), p. 816–(Serie A. Normas e Manuais Tecnicos)

[3] Brasil. Ministerio da Saude: ' Guia de leptospirose: diagnostico e manejo clinico'. Ministerio da Saude, Secretaria de Vigilancia em Saude. Departamento de Vigilancia das Doencas Transmissiveis' (Ministerio da Saude, Brasilia, 2014)

[4] García-Laencina, P.J., Abreu, P.H., Abreu, M.H*., et al.*: 'Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values', *Comput. Biol. Med.*, 2015, **59**, pp. 125–133

[5] Yeh, J.Y., Wu, T.H., Tsao, C.W.: 'Using data mining techniques to predict hospitalization of hemodialysis patients', *Decis. Support Syst.*, 2011, **50**, (2), pp. 439–448

[6] Yeh, D.Y., Cheng, C.H., Chen, Y.W.: 'A predictive model for cerebrovascular disease using data mining', *Expert Syst. Appl.*, 2011, **38**, (7), pp. 8970–8977

[7] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: 'From data mining to knowledge discovery in databases', *AI Mag.*, 1996, **17**, (3), p. 37

[8] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: 'The KDD process for extracting useful knowledge from volumes of data', *Commun. ACM*, 1996, **39**, (11), pp. 27–34

[9] Sahle, G., Meshesha, M.: 'Uncovering knowledge that supports malaria prevention and control intervention program in Ethiopia', *Electron. J. Health Inform.*, 2013, **8**, (1), p. 7

[10] Bakar, A.A., Kefli, Z., Abdullah, S*., et al.*: 'Predictive models for dengue outbreak using multiple rule base classifiers'. Electrical Engineering and Informatics (ICEEI), 2011 Int. Conf. on IEEE, 2011, pp. 1–6

[11] Bier, D., Molento, M.B.: 'Distribuicao espacial e fatores de risco para leptospirose canina na Vila Pantanal, Curitiba, Parana, Brasil'. *ScM thesis*, Universidade Federal do Praná, 2012

[12] da Costa Côrtes, S., Porcaro, R.M., Lifschitz, S.: 'Mineração de dados-Funcionalidades, técnicas e abordagens'. PUC, 2002

[13] Oguntimilehin, A., Adetunmbi, A.O., Abiola, O.B.: 'A review of predictive models on diagnosis and treatment of malaria fever', *Int. J. Comput. Sci. Mobile Comput.*, 2015, **4**, pp. 1087–1093

[14] Ugwu, C., Onyejegbu, N.L., Obagbuwa, I.C.: 'The application of machine learning technique for malaria diagnosis'. Nigeria Computer Society 23rd National Conf., 2013, pp. 151–158

[15] Hall, M., Frank, E., Holmes, G*., et al.*: 'The WEKA data mining software: an update', *ACM SIGKDD Explorations Newsletter*, 2009, **11**, (1), pp. 10–18

[16] Rocha, N.Jr, Nivison, R., Barreiro, C*., et al.*: '*Classification model analysis for the prediction of leptospirosis cases*'. Actas de la 11^a Conferencia Ibérica de Sistemas y Tecnologías de Información, Gran Canaria, España, June 2016, pp. 966–971

[17] Ko, A.I., Reis, M.G., Dourado, C.M.R*., et al.* , Salvador Leptospirosis Study Group: 'Urban epidemic of severe leptospirosis in Brazil', *The Lancet*, 1999, **354**, (9181), pp. 820–825

[18] Araujo, Wildo Navegantes de and Reis, Mitermayer Galvão: 'Aspectos epidemiológicos da leptospirose no Brasil, 2000 a 2009 e a avaliação do conhecimento e das atitudes sobre a doença em uma favela na cidade de Salvador, Bahia. 114 f.'. *PhD thesis, Fundação Oswaldo Cruz*, Centro de Pesquisas Gonçalo Moniz, Salvador, 2010

[19] Frank, E., Hall, M., Holmes, G*., et al.*: 'Weka-a machine learning workbench for data mining'. Data mining and knowledge discovery handbook, Springer USA, 2009, pp. 1269–1277

[20] Baldi, P., Brunak, S., Chauvin, Y*., et al.*: 'Assessing the accuracy of prediction algorithms for classification: an overview', *Bioinformatics*, 16.5, 2000, **16**, (5), pp. 412–424

[21] Lalkhen, A.G., McCluskey, A.: 'Clinical tests: sensitivity and specificity', *Contin. Educ. Anaesth. Crit. Care Pain*, 2008, **8**, (6), pp. 221–223

[22] Costa, F., Hagan, J.E., Calcagno, J*., et al.*: 'Global morbidity and mortality of leptospirosis: a systematic review', *PLoS Negl. Trop. Dis.*, 2015, **9**, (9), p. e0003898

[23] Bergmeir, C., Hyndman, R.J., Koo, B.: 'A note on the validity of cross-validation for evaluating time series prediction'. Monash University Department of Econometrics and Business Statistics Working Paper, 10, 15, 2015

[24] Dean, A.G., Arner, T.G., Sangam, S*., et al.*: '*Epi Info 2002, a database and statistics program for public health professionals for use on Windows 95, 98, ME, NT, 2000 and XP computers*' (Centers for Disease Control, Atlanta, 2002)