Journal of Information Technologies, Vol. 7, No. 2, 2014, ISSN: 1337-7467

Diagnosis of diabetes for Pima Indian female patients with a propositional rule learner

Miroslav Benedikovič, Department of Informatics at the University of Žilina, Slovakia

Abstract: A propositional rule learner using collected medical data is analysed for the purposes of diagnosis of diabetes for Pima Indian female patients who are at least 21 years old. The quality of the diagnosis is judged with the aim of minimization of life-threatening situations and costs. Achieved experimental results regarding the performance of the prepared system for diagnosis are provided and they are also compared with a C4.5 decision tree.

Key words: propositional rule learner, classification, confusion matrix, diabetes, Pima Indian female patients

1. Introduction

One of the most important issues for healthcare organizations is provision of quality services at affordable costs (3). Quality service implies diagnosing patients correctly and administering treatments that are cost-effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Healthcare organizations must also minimize the cost of clinical tests. The use of data mining techniques can be helpful for assistance in medical diagnosis as it can be seen in (3)(4)(5)(10).

Specifically, a propositional rule learner is used for diagnosis of diabetes among the Pima Indian females so that a general opinion of this technique is obtained and future analysis can be possible. The female Pima Indian patients who are at least 21 years old as these ones have 19 times higher diabetic rate than a typical person in Minnesota (8). Previously known recent studies suggest that a propositional rule learner can be effective to classify medical data (6). The main aim here is its classification accuracy targeted at minimization of life-threatening situations and minimization of costs in diagnosis of diabetes for the Pima Indian females and it is compared with a previous study of the author (2).

The organization of the paper is as follows. The diabetes data about Pima Indians used for the experiments with the algorithm presented in Section 3 is described in Section 2. The results of experiments are shown in Section 4 and the conclusions are presented in Section 5.

2. Diabetes data about Pima Indians

The diabetes data was obtained from the UCI Repository and originates in the National Institute of Diabetes and Digestive and Kidney Diseases (1)(9). It has eight numerical attributes classified into a positive/negative class for diabetes and they are described in Table. 1. There are 500 negative patients and 268 positive patients and the data describes 768 Pima Indian females living near Phoenix, Arizona, USA. Pima Indian females have a large amount of patients suffering from diabetes according to the World Health Organization.

Table. 1: Pima Indians Diabetes Database.

Attribute	**Type**	**Description**	**Unit**
A_1	Numerical	Diastolic blood pressure.	mm Hg
A_2	Numerical	Age.	years
A_3	Numerical	2-hour serum insulin.	mu U/ml
A_4	Numerical	Diabetes pedigree function.	N/A
A_5	Numerical	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.	mmol/L
A_6	Numerical	Number of times pregnant.	No.
A_7	Numerical	Triceps skin fold thickness.	mm
A_8	Numerical	Body mass index.	kg/m^2
C	Categorical	Whether the patient shows signs of diabetes according to World Health Organization criteria. Categorical values:	N/A

		positive - positive test for diabetes, *negative* – negative test for diabetes.	

3. Propositional rule learner

The propositional rule learner used is based on the algorithm described in (6) and introduced in (7). It uses a set of rules R combined together with current values of attributes in $\boldsymbol{A}$ for a patient $\boldsymbol{p}$, i.e. all $A_k(\boldsymbol{p})$, $A_k \in \boldsymbol{A}$. The set of rules R is created once and then this set is employed for prognosis for a longer period of time. The set of rules is created as follows:

Set initially $R = \emptyset$ and for both $c_j \in C$ from the less prevalent one to the more frequent one, execute:

1. Building stage:

Repeat 1.1 and 1.2 until the descrition length of R and patients $\boldsymbol{p} \in \boldsymbol{P}$ is 64 bits greater than the smallest descrition length met so far, or there are no positive patients, or the error rate >= 50%.

1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is as accurate as possible. Every possible value of each attribute is tried and the condition with highest information.

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents; The pruning metric is (p-n)/(p+n) where p is the number of c_2 and n is the number of c_1. However, it is actually 2p/(p+n) - 1, so in the implementation the following is used: (p+1)/(p+n+2). Thus if p+n is 0, it is 0.5.

2. Optimization stage:

After generating the initial rules $R_i \in R$, generate and prune two variants of each rule R_i from randomized data using procedure 1.1 and 1.2. But one

variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (tp+tn)/(p+n) where tp is the number of true positives and tn is the number of true negatives. Then the smallest possible descrition length for each variant and the original rule is computed. The variant with the minimal descrition length is selected as the final representative of R_i in R. After all $R_i \in R$ have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from R that would increase the descrition length of the whole R if it were in it. Distributions of values in C are also computed for each rule in R.

Prediction for a particular patient $\boldsymbol{p}$ is as follows:

For each rule r in R execute:

If r covers values $A_k(\boldsymbol{p})$, $A_k \in \boldsymbol{A}$, value $c_j \in C$ with the maximal value in the distribution related to rule r is chosen as the result. STOP.

4. Experiments

Conduction of experiments was carried out with tool Weka - Waikato Environment for Knowledge Analysis (11). The quality of the diagnosis was measured with sensitivity = true positives / (true positives + false negatives), specificity = true negatives / (false positives + true negatives), positive predictive value = true positives / (true positives + false positives), negative predictive value = true negatives / (true negatives + false negatives), and accuracy = (true positives + true negatives) / (true positives + true negatives + false positives + false negatives). True positives, true negatives, false positives, and false negatives were computed in 10-fold cross-validation. There should not be negative diagnosis for patients with diabetes as this could lead to life-threatening situations (measured by sensitivity). At the same time, there should not be a large number of patients labelled as positive if they are negative because this would increase the running costs of the decision support system (measured by specificity). As a consequence, the sum of sensitivity and specificity measures both the life-threatening situations and the running costs of the system where higher values are preferred.

Table. 2: Experimental results.

Measure	**Propositional rule learner**	**C4.5 decision tree**
Sensitivity	0.6442687747	0.63241106719
Specificity	0.79611650485	0.79029126213
Sensitivity + Specificity	1.44038527955	1.42270232932
Positive Predictive Value	0.60820895522	0.59701492537
Negative Predictive Value	0.82000000000	0.81400000000
Accuracy	0.74609375000	0.73828125000

The propositional rule learner is implemented in Weka as class JRip and the C4.5 decision tree is implemented in Weka as class J48. The results of conducted experiments are presented in Table. 2 where sensitivity for the propositional rule learner is 0.6442687747, specificity is 0.79611650485, and the sum of sensitivity and specificity is 1.44038527955. The experiments with the C4.5 decision tree were conducted within the study of (2). The results of the propositional rule learner are slightly better than the results of the C4.5 decision tree.

5. Conclusions

A propositional rule learner was employed for diagnosis of Pima Indian female patients. The data had 768 patients described by eight numerical attributes. The accuracy of the propositional rule learner was evaluated in 10-fold cross-validation using sensitivity, specificity, positive predictive value, and negative predictive value. As minimization of life-threatening situations and minimization of running costs was targeted, the sum of sensitivity and specificity was computed as sensitivity is associated with life-threatening situations and specificity is associated with running costs. The following values were obtained. Sensitivity was 0.63241106719, specificity was 0.79029126213, positive predictive value was 0.59701492537, negative predictive value was 0.81400000000, and the sum was 1.42270232932. The sum was slightly better than the sum of sensitivity and specificity for the J48 decision tree.

6. Bibliography

(1) Bache, K., Lichman, M. (2013) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

(2) Benedikovic, M. (2014) Decision tree applied to diagnosis of diabetes for Pima Indian female patients, Journal of Information Technologies 7(1): 10-17.

(3) Bohacik, J., Davis, D. N. (2010) Data mining applied to cardiovascular data, Journal of Information Technologies 3(2): 14-21.

(4) Bohacik, J., Davis, D. N. (2012) Diagnosis and management of cardiovascular disease with an intelligent decision-making support system, ULAB Journal of Science and Engineering 3(1): 2-6.

(5) Bohacik, J., Kambhampati, C., Davis, D. N., Cleland, J. (2013) Prediction of mortality rates in heart failure patients with data mining methods, Annales UMCS, Informatica 13(2): 7-16.

(6) Bohacik, J., Kambhampati, C., Davis, D. N., Cleland, J. (2014) Use of a propositional rule learner for prognosis of mortality rates in heart failure patients, Journal of Information Technologies 7(1): 1-9.

(7) Cohen W. W. (1995) Fast Effective Rule Induction. Proceedings of the 20th International Conference on Machine Learning, pp. 115-123.

(8) Knowler, W. C., Bennett, P. H. et al. (1978) Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. American Journal of Epidemiol 108(6): 497-505.

(9) Mao, Y., Chen, Y., Hackmann, G., Chen, M., Lu, C., Kollef, M., Bailey, T. C. (2011) Medical data mining for early deterioration warning in general hospital wards. Proceedings on the IEEE 11th International Conference on Data Mining Workshops, pp. 1042 - 1049.

(10) Trnka, A., Kovarova, M., Doci, I. (2011) Occurrence of sexual disorders in Slovakia - a five-year analysis with Data Mining, Forum Staticum Slovacum 7(2): 185-192.

(11) Witten I. H., Frank E., Hall M. A. (2011) Practical machine learning tools and techniques (Third Edition). USA: Morgan Kaufman.

7. Address of the author:

RNDr Miroslav Benedikovič
Department of Informatics
Faculty of Management Science and Informatics
University of Žilina
010 26 Žilina
SLOVAKIA
Miroslav.Benedikovic@fri.uniza.sk

The article was accepted for publication in October 2014 by the publisher of the Journal of Information Technologies (Vol. 7, No. 2, ISSN: 1337-7467), i.e. by the Faculty of Mass Media Communication, University of Ss. Cyril and Methodius in Trnava, Slovakia.