Decision tree applied to diagnosis of diabetes for Pima Indian female patients

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

Abstract: Large amounts of data in hospitals lead to development of medical decision support systems for hospitals so that diagnostic accuracy and costs are targeted. In this paper, a C4.5 decision tree 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.

Key words: decision tree, classification, confusion matrix, diabetes

1. Introduction

Medical data mining is becoming a more and more used method for diagnosis as it can be seen in the number of various recent papers published in this area (2)(11)(3)(4)(5)(6)(8)(9)(10)(17)(20)(23)(24)(7)(18)(21)(12)(14)(15)(22)(16). The data mined there is somewhat unique and may impose constrains and difficulties which are related to uncertainties, heterogeneity, missing values, volume, privacy, and so on. Another issue is that medicine has a special status as the outcomes of medical care are life or death. In addition to it, medicine is necessary and is not just an optional luxury. For these reasons, there is an urgent need for solutions and work by data mining and medical experts and society is prepared to participate by allocations resources.

The specific work presented here is aimed at 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 (13). Precisely, a C4.5 decision tree 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. Previously known recent studies suggest that a C4.5 decision tree can be quick to make and effective to classify medical data (3). On the other hand, its interpretability may be somewhat limited as discretization of numerical attributes is done automatically and without a

clinician's opinion or it can have sharp boundaries which can cause unpredictable diagnosis. However, 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.

The paper is organized as follows. Section 2 describes the Pima Indians Diabetes Database used for the experiments with the algorithm presented in Section 3. Section 4 shows the results of experiments. Section 5 concludes the paper.

2. Pima Indians Diabetes Database

The data about female Pima Indian patients originally comes from the National Institute of Diabetes and Digestive and Kidney Diseases, USA and it was obtained from the UCI Repository (1)(19). The data is diagnostic with eight numerical variables and one binary class variable and is whether the Pima Indian female patient living near Phoenix, Arizona, USA shows signs of diabetes according to the World Health Organization. The data contains information about 768 patients whose attributes are described in Table. 1. Class distribution is 500 negative patients and 268 positive patients.

Table. 1: Pima Indians Diabetes Database.

Attribute	**Type**	**Description**
A_1	Numerical	Number of times pregnant.
A_2	Numerical	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
A_3	Numerical	Diastolic blood pressure (mm Hg).
A_4	Numerical	Triceps skin fold thickness (mm).
A_5	Numerical	2-Hour serum insulin (mu U/ml).
A_6	Numerical	Body mass index.
A_7	Numerical	Diabetes pedigree function.
A_8	Numerical	Age (years).

C	Categorical	Whether the patient shows signs of diabetes according to World Health Organization criteria. Categorical values: *positive* - positive test for diabetes, *negative* – negative test for diabetes.

3. C4.5 decision tree

The decision tree method uses a C4.5 decision tree created on the basis of the Pima Indians Diabetes Database containing 768 patients described in the previous section. The tree consists of the following nodes (4): a) the root associated with an attribute A_k; b) internal nodes associated with an A_k; c) leaf nodes associated with a categorical value defined for *C* (*positive* or *negative*). At each node of the tree (except the leaf nodes), the attribute of the data that most effectively splits its set of patients into subsets enriched in *positive* and *negative* is chosen. The normalised information gain is used as the splitting criterion while the attribute with its highest value is chosen to make the decision. Each non-leaf node has an outgoing branch for each possible categorical value of $A_k$where these categorical values are created automatically through discretization. The decision tree is pruned after creation so that some branches are removed, which can improve the prediction power of the tree (especially when all female Pima Indian patients are taken into consideration). Each non-leaf node in the decision tree represents an attribute A_k for a patient to be diagnosed and each branch represents a categorical value that the node can assume.

Diagnosis of a particular female Pima Indians patient is as follows:

Start at the root node, label the root node as current node, and do the following:

Look at the values associated with all outgoing branches of the current node. Move to do node which is on the other side of the outgoing branch with the value that matches the value given about the patient. Label the node where you just moved as the current node. If this node is a leaf node, give the value associated with it as diagnosis. Otherwise, repeat.

4. Experiments

The experiments were conducted using the Waikato Environment for Knowledge Analysis – Weka (25), version 3.6.11, which is a tool developed by the University of Waikato, New

Zealand. The C4.5 algorithm is implemented as class J48. 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 regarding the C4.5 decision tree.

Measure	**Value**
Sensitivity	0.63241106719
Specificity	0.79029126213
Sensitivity + Specificity	1.42270232932
Positive Predictive Value	0.59701492537
Negative Predictive Value	0.81400000000
Accuracy	0.73828125000

The results of conducted experiments with the decision tree are presented in Table. 2 where sensitivity is 0.63241106719, specificity is 0.79029126213, and the sum of sensitivity and specificity is 1.42270232932.

5. Conclusions

A decision support system with a C4.5 decision tree was employed for diagnosis of Pima Indian female patients. The data had 768 patients described by eight attributes. The accuracty of the decision tree 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 best value of the sum is 2.0 and the worst value of the sum is 0.0. 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. A higher value of the sum of sensitivity and specificity could possibly be achieved by incorporation of uncertainties and notions of fuzzy logic into the C4.5 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) Bohacik, J., Kambhampati, C., Davis, D. N., Cleland, J. (2014) Use of cumulative information estimations for risk assessment of heart failure patients. Proceedings of the IEEE International Conference on Fuzzy Systems - FUZZ-IEEE 2014, Special Session: Fuzzy Decision Making and Decision Support Systems I - WeF2-4, pp. 1402-1407.

(3) Bohacik, J., Kambhampati, C., Davis, D. N., Cleland, J. (2013) Analysis of fuzzy decision trees on expert fuzzified heart failure data, Proc. of the IEEE International Conference on Systems, Man, and Cybernetics - IEEE SMC, Special Session: Soft Computing - C12-01.

(4) 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.

(5) Bohacik, J., Kambhampati, C., Davis, D. N., Cleland, J. (2013) Alternating decision tree applied to risk assessment of heart failure patients, Journal of Information Technologies 6(2): 25-33.

(6) Bohacik, J., Davis, D. N. (2013) Fuzzy rule-based system applied to risk estimation of cardiovascular patients, Journal of Multiple-Valued Logic and Soft Computing 20(5-6): 445-466.

(7) 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.

(8) Cazzolato, M. T., Ribeiro, M. X. (2013) A statistical decision tree algorithm for medical data stream mining. IEEE 26th International Symposium on Computer-Based Medical Systems, pp. 389 - 392.

(9) Hutterer, S., Mayr, S., Zauner, G., Silye, R. (2013) Data mining supported analysis of medical atomic force microscopy images. Proceedings of the IEEE Symposium on Computational Intelligence in Healthcare and e-health, pp. 99-104.

(10) Ilayaraja, M., Meyyappan, T. (2013) Mining medical data to identify frequent diseases using Apriori algorithm. Proceedings of the 2013 International Conference on Pattern Recognition, pp. 194-199.

(11) Martti Juhola, Youming Zhang, Jyrki Rasku (2013) Biometric verification of a subject through eye movements. Computers in Biology and Medicine 43(1): 42-50.

(12) Khaing, H. W. (2011) Data mining based fragmentation and prediction of medical data. Proceedings of the 3rd International Conference on Computer Research and Development, pp. 480-485.

(13) 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.

(14) Lao, Y., Gu, Q., Liang, Z., Tan, D. (2011) A data mining research method based on the concept of evidence based TCM inheritance in famous veteran TCM doctors' personal medical records. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp. 746-748.

(15) 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.

(16) Miettinen, K., Juhola, M. (2010) Classification of Otoneurological Cases According to Bayesian Probabilistic Models. Journal of Medical Systems 34(2): 119-130.

(17) Ranganatha, S., Pooja Raj, H.J., Anusha, C., Vinay, S.K. (2013) Medical data mining and analysis for heart disease dataset using classification techniques. Proceedings of the National Conference on Challenges in Research & Technology in the Coming Decades, pp. 1-5.

(18) Robu, R., Hora, C. (2012) Medical data mining with extended WEKA. Proceedings of the IEEE 16th International Conference on Intelligent Engineering Systems, pp. 347 - 350.

(19) Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Symposium on Computer Applications and Medical Care, pp. 261-265, IEEE Computer Society Press.

(20) Sumalatha, G., Muniraj, N.J.R. (2013) Survey on medical diagnosis using data mining techniques. Proceedings of the International Conference on Optical Imaging Sensor and Security, pp. 1-8.

(21) Titapiccolo, J.I., Ferrario, M., Cerutti, S., Signorini, M.G. (2012) Mining medical data to develop clinical decision making tools in hemodialysis. Proceedings of the IEEE 12th International Conference on Data Mining Workshops, pp. 99-106.

(22) 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.

(23) Ullrich, M., ten Hagen, K. ; Lassig, J. (2013) A data mining approach to reduce the number of maintenance visits in the medical domain. IEEE 7th International

Conference on Intelligent Data Acquisition and Advanced Computing Systems, pp. 255-258.

(24) Wagle, S., Mangai, J.A., Kumar, V.S. (2013) An improved medical image classification model using data mining techniques. Proceedings of the 7th IEEE GCC Conference and Exhibition, pp. 114-118.

(25) 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 August 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.