A COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MULTILABEL CLASSIFICATION OF CAN

A.V. KELAREV, A. STRANIERI, J.L. YEARWOOD

Centre for Informatics and Applied Optimization, University of Ballarat, P.O. Box 663, Ballarat, Victoria 3353, Australia
Email: {a.kelarev,a.stranieri,j.yearwood}@ballarat.edu.au

H.F. JELINEK

Centre for Research in Complex Systems and School of Community Health, Charles Sturt University, PO Box 789, Albury, NSW 2640, Australia
Email: hjelinek@csu.edu.au

Abstract

This article is devoted to the investigation and comparison of several important machine learning algorithms in their ability to obtain multi-label classifications of the stages of cardiac autonomic neuropathy (CAN). Data was collected by the Diabetes Complications Screening Research Initiative at Charles Sturt University. Our experiments have achieved better results than those published previously in the literature for similar CAN identification tasks.

Machine learning methods and automated data mining are important for health informatics and have been actively investigated, for example, in [2], [4], [5], [6], [7] and [9]. The present article is devoted to an experimental comparison of several important classification methods in their ability to obtain multi-label classifications of cardiac autonomic neuropathy of an extensive data set collected by the Diabetes Complications Screening Research Initiative (DiScRi) at Charles Sturt University, [3].

2010 Mathematics Subject Classification: 68T01, 68T10
Keywords and phrases: machine learning, multilabel classification, cardiac autonomic neuropathy.

Cardiac autonomic neuropathy is a condition associated with damage to the autonomic nervous system innervating the heart and highly prevalent in people with diabetes, [1], [3]. It is known as one of the causes of mortality among type 2 diabetes patients. The identification of CAN has been investigated previously in the literature, see [2], but due to lack of suitable data sets multi-label classification of various stages of CAN has not been considered. The classification of disease progression associated with CAN is important, because it has implications for planning of timely treatment, which can lead to an improved well-being of the patients and a reduction in morbidity and mortality associated with cardiac arrhythmias in diabetes.

This paper utilizes the clinical test results and health-related demographic parameters collected at the Diabetes Complications Screening Research Initiative, DiScRi, organised at Charles Sturt University, [3], [4], [5]. There are no alternative data sets containing comparable collections of comprehensive test outcomes related to CAN.

A preprocessing system has been implemented in Python to automate several expert editing rules that can be used to reduce the number of missing values in the data base. Preprocessing of data produced 1299 complete rows with complete values and 200 features, which were used for the experimental evaluation of the performance of data mining algorithms. We have created several lists ranking the features in the order of their relevance to the multi-label classification of CAN and used them in consultation with the experts maintaining the data base to select the most essential features.

Several of the most important machine learning classifiers were trained to recognize four classes of CAN given in the data base (see Table 1). Please refer to [10] for background information and preliminaries (see also [6], [7] and [11]). Ten-fold cross validation was applied to determine the effectiveness of these classifiers. For the same features VAHR, DBHR, HGBP and QRS used in [2], see also [1], our experimental results are presented in Table 1.

	Accuracy	Precision	Recall	ROC area
BayesNet	70.05	0.686	0.701	0.866
DecisionTable	76.98	0.781	0.770	0.899
FURIA	85.07	0.856	0.851	0.917
J48	88.30	0.885	0.883	0.937
JRip	81.83	0.818	0.818	0.881
NaiveBayes	54.58	0.575	0.546	0.732
PART	88.99	0.891	0.890	0.933
RandomForest	90.84	0.910	0.908	0.908
RBFNetwork	64.28	0.603	0.643	0.789
Ridor	83.53	0.836	0.835	0.871
SMO	60.74	0.496	0.607	0.661

Table 1. Multi-label classifications for stages of CAN

The best results were obtained using Random Forest, with higher performance compared to previous outcomes in [2]. This improvement is even more significant, because [2] considered a much simpler binary classification, selected 291 rows, and did not use ten-fold cross validation. The outcomes of the present paper are very good for classification of CAN, and also when compared to recent results obtained for other data sets using different methods, for example, in [2], [3], [4] and [5].

References

[1]D.J. Ewing, Diabetic autonomic neuropathy and the heart, Diabetes Research and Clinical Practice 30 (1996), S31-S36.

[2]S. Huda, H. Jelinek, B. Ray, A. Stranieri and J. Yearwood, Exploring novel features and decision rules to identify cardiovascular autonomic neuropathy using a hybrid of wrapper-filter based feature selection, ISSNIP 2010, Intelligent Sensors, Sensor Networks & Information Processing, 297-302.

[3]H.F. Jelinek, C. Wilding, P. Tinley, An innovative multi-disciplinary diabetes complications screening programme in a rural community: A

description and preliminary results of the screening, Australian Journal of Primary Health 12 (2006), 14-20.

[4]H.F. Jelinek, A. Khandoker, M. Palaniswami and S. McDonald, Heart rate variability and QT dispersion in a cohort of diabetes patients, Computing in Cardiology 37 (2010), 613-616.

[5]H. Jelinek, A. Rocha, T. Carvalho, S. Goldenstein and J. Wainer, Machine learning and pattern classification in identification of indigenous retinal pathology, Proc. IEEE Conf. Eng. Med. Biol. Soc. (2011), 5951-5954.

[6]B. Kang, A. Kelarev, A. Sale and R. Williams, A new model for classifying DNA code inspired by neural networks and FSA, Lecture Notes in Computer Science 4303 (2006), 187-198.

[7]A.V. Kelarev, An algorithm for repeated convex regions in geographic information systems, Far East J. Appl. Math. 8 (2002), 75-79.

[8]A.V. Kelarev, Computing statistics for polynomial codes: an algorithm based on the Mann-Whitney U-test, Advances & Applications in Statistics 6 (2006), 53-56.

[9]A. Kelarev, B. Kang and D. Steane, Clustering algorithms for ITS sequence data with alignment metrics, Lecture Notes in Artificial Intelligence 4304 (2006), 1027-1031.

[10]I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/Morgan Kaufman, Amsterdam, 2005.

[11]J. Yearwood, D. Webb, L. Ma, P. Vamplew, B. Ofoghi and A. Kelarev, Applying clustering and ensemble clustering approaches to phishing profiling, Data Mining and Analytics 2009, Proc. 8th Australasian Data Mining Conference, AusDM 2009, Current Research and Practice in Information Technology 101 (2010), 25-34.