Advances in Computer Science and Engineering

Published Online: November 2014
Available online at http://pphmj.com/journals/acse.htm
Volume 13, Number 1, 2014, Pages 69-72

ENSEMBLES OF BINARY DECISION TREES FOR PREDICTING AIR QUALITY

H. F. Jelinek[1], A. V. Kelarev[1], A. Kolbe[2], S. Heidenreich[3] and T. Oakman[4]

[1]Centre for Research in Complex Systems and School of Community Health
Charles Sturt University
P. O. Box 789, Albury, NSW 2640, Australia
e-mail: hjelinek@csu.edu.au
andreikelarev-charlessturtuniversity@yahoo.com

[2]Faculty of Science
Charles Sturt University
Wagga Wagga, Locked Bag 588, NSW 2678, Australia
e-mail: akolbe@csu.edu.au

[3]New South Wales Office of Environment and Heritage
Gunnedah Research Centre
Gunnedah, P. O. 20, NSW 2380, Australia
e-mail: stephan.heidenreich@environment.nsw.gov.au

[4]New South Wales Public Health Unit
P. O. Box 3095, Albury, NSW 2640, Australia
e-mail: tracey.oakman@gsahs.health.nsw.gov.au

Abstract

This paper concentrates on applying a novel data mining algorithm, the Ensemble of Binary Decision Trees, EBDT, for the detection and monitoring of environmental particulate matter in high-risk areas due

Received: June 16, 2014; Accepted: July 14, 2014
Keywords and phrases: machine learning, multilabel classification, air pollution.

to agricultural stubble burning. Experimental outcomes presented here show that the EBDT classifier based on J48 achieved the best outcome for the detection of PM2.5 patterns with an accuracy of 83.70%.

Data mining is used across many sectors including agriculture and health (cf. [1] to [7]). This article introduces a novel method, called Ensemble of Binary Decision Trees, EBDT ensemble classifier. Our experiments investigate the effectiveness of the EBDT classifiers for the prediction of the NSW classification of the level of PM2.5 particles in the air.

Particulate matter (PM) in the air is linked with a range of health effects manifesting as increased rates of both morbidity and mortality, [7]. It is an important part of air quality assessment. The depth of penetration of PM into the respiratory tract depends on the size of the particles with fine particles penetrating more deeply into the lungs. Fine and ultrafine particles play a dominant role in human health impacts [7]. A commonly described PM fraction referring to the particles with an aerodynamic diameter of at most 2.5μm is denoted by PM2.5.

In New South Wales (NSW), air pollution is monitored by the NSW Office of Environment and Heritage (OEH). OEH expresses air pollution data in terms of an Air Quality Index (AQI) with six categories of air quality defined in terms of all categories of particular matter as follows: very good (up to 33μg), good (from 34μg to 66μg), fair (from 67μg to 99μg), poor (from 100μg to 149μg), very poor (from 150μg to 199μg), hazardous (200μg and higher). Here we apply the same categories to the amount of PM2.5 in the air.

We introduce a novel method for air quality prediction, called an Ensemble of Binary Decision Trees, EBDT. In this method we train a collection of binary decision trees, one for each of the thresholds. Each tree is trained to recognise two classes: the class of instances with the PM2.5 value less than the threshold, and the class of instances with the PM2.5 value greater than the threshold. After all binary decision trees have been trained,

they operate as an ensemble in the following fashion. To recognise instances of each class, two binary decision trees corresponding to the upper and lower threshold of the class operate in conjunction. A particular instance is recognised as belonging to the class if the binary decision tree corresponding to the upper threshold confirms that the class value of the instance does not exceed the top threshold and at the same time to binary decision tree corresponding to the lower threshold confirms that the class value of the instance is greater than the lower threshold. Thus, an ensemble of 6 binary classifiers is used to recognise all classes.

We used a large data set for Wagga Wagga obtained from the DustWatch program ([1], [8]) and the Bureau of Meteorology, Australia over a twelve months period (1 August 2011-31 July 2012), currently consisting of 18,485 records, which is a large dataset. We considered the problem of predicting the NSW classes for the level of PM2.5 in the air.

Our experiments investigated and compared EBDT classifiers based on the following decision trees: J48, NBTree, RandomTree, REPTree, SimpleCart. Let us refer to [2] and [9] for background information on these algorithms. In order to avoid overfitting, we used the standard ten-fold cross validation (cf. [9]). Ten-fold cross validation was applied to determine the effectiveness of these regression methods. Our experimental results are presented in the following table:

Regression Methods	Accuracy (%)	
	Trees	EBDT ensembles
ADTree	73.84	80.59
J48	76.16	83.70
NBTree	75.38	82.23
RandomTree	73.13	80.40
REPTree	73.83	81.24
SimpleCart	75.21	82.13

The best results were obtained using EBDT ensemble based on J48 tree. This is an excellent outcome, since we considered all six of the NSW classes for PM2.5, which has not been investigated before.

References

[1] M. C. Baddock, C. L. Strong, J. F. Leys, S. K. Heidenreich, E. K. Tews and G. H. McTainsh, A visibility and total suspended dust relationship, Atmospheric Environment 89 (2014), 329-336.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, The WEKA data mining software: an update, SIGKDD Explorations 11 (2009), 10-18.

[3] H. F. Jelinek, A. V. Kelarev, A. Stranieri and J. L. Yearwood, Rule-based classifiers and meta classifiers for identification of cardiac autonomic neuropathy progression, Int. J. Information Science and Computer Mathematics 5 (2012), 49-53.

[4] A. V. Kelarev, A. Stranieri, J. L. Yearwood and H. F. Jelinek, A comparison of machine learning algorithms for multilabel classification of CAN, Advances in Computer Science and Engineering 9 (2012), 1-4.

[5] A. V. Kelarev, An algorithm for repeated convex regions in geographic information systems, Far East J. Appl. Math. 8(1) (2002), 75-79.

[6] A. V. Kelarev, Computing statistics for polynomial codes: an algorithm based on the Mann-Whitney U-test, Adv. Appl Stat. 6(1) (2006), 53-56.

[7] S. R. Kirkhorn and V. F. Garry, Agricultural lung diseases, Environmental Health Perspectives 108 (2000), 705-712.

[8] J. Leys, G. McTainsh, C. Strong, S. Heidenreich and K. Biesaga, DustWatch: using community networks to improve wind erosion monitoring in Australia, Earth Surface Processes and Landforms 33 (2008), 1912-1926.

[9] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/Morgan Kaufman, Amsterdam, 2011.