JNCC2: an extension of naive Bayes classifier suited for small and incomplete data sets

Giorgio Corani, Marco Zaffalon

IDSIA
Galleria 2, CH-6928 Manno (Lugano)
Switzerland
{*giorgio,zaffalon*}*@idsia.ch*

Abstract

JNCC2 implements the Naive Credal Classifier 2 (NCC2), i.e., an extension of naive Bayes to imprecise probabilities, designed to return robust classification even on small and/or incomplete data sets, which is often the case in environmental case studies.

Key words: classification, naive credal classifier 2, naive bayes

Software availability

Name: JNCC2.
Developers: Giorgio Corani and Marco Zaffalon.
Affiliation: IDSIA (`www.idsia.ch`), Manno, Switzerland.
License: GNU GPL (open source).
Availability: executable, sources and manuals downloadable from `www.idsia.ch/~giorgio/jncc2.html` .
Year first available: 2007.
Software required: Java Runtime Environment 5.0 or higher.
Programming language: Java.
Operating system: OS independent.
User interface: command-line.

1 Overview

Classifiers learn from data the relationship that holds between a set of attributes characterizing a given object, and the class of the object. In environmental modelling, classifiers are used, for instance, to predict the presence or absence of a species on a certain area, given some features of the landscape and of the soil (e.g., Garzon et al. (2006)), or to analyze remotely sensed images to identify the land cover type of each pixel (e.g., Brown et al. (1999)). According to Nunez et al. (2004), the reliability of the environmental decision support system could be improved by incorporating classification algorithms, besides more traditional modelling tools.

A simple yet effective approach for classification is naive Bayes (Duda and Hart, 1973). The Naive Credal Classifier 2 (Corani and Zaffalon, 2007b) generalizes naive Bayes towards imprecise probabilities, with the goal of providing higher reliability. When faced with instances that are hard to classify, NCC2 returns *set-valued* classifications; for instance, if the possible classes for the pixels of a remotely sensed image are 'crops', 'grassland' and 'water', the output might be both 'crops' and 'grassland'. Imprecise classifications arise especially when the data sets are small and/or incomplete, which is often the case in environmental problems. The experimental analysis of Corani and Zaffalon (2007b) reports that the accuracy of naive Bayes is quite low on the instances which are classified in a set-valued way by NCC2, and hence that NCC2 outputs more than one class on truly doubtful instances. NCC2 is an extensions, characterized by more advanced algorithms for the treatment of missing data, of the former naive credal classifier, which had been used for instance by Zaffalon (2005) to predict the abundance of grass grub as a function of biotic factors and farming practices.

JNCC2 is the Java implementation of NCC2; it is hence platform-independent. Despite being implemented in an interpreted language, JNCC2 is fast: training and validating the classifier takes only a few seconds, even if the data set contains several thousands of instances. In fact, NCC2 algorithms, deriving from naive Bayes, are computationally efficient.

JNCC2 learns both naive Bayes and NCC2 on the testing set, and then validates them on a subset of data different from the training set. Validation can be performed only once (single testing set) or several times (cross-validation). JNCC2 reports then to file the performance indicators for both naive Bayes and NCC2.

JNCC2 loads data from ARFF files; this is a plain text format, developed for WEKA (Witten and Frank, 2005), an open-source software for data mining. Large public repositories of ARFF data sets are available, starting from the WEKA webpage of data sets (`http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html`).

JNCC2 runs from the command-line; the user manual (Corani and Zaffalon, 2007a), downloadable from the website, reviews all the available options and illustrates several worked examples.

References

Brown, M., Gunn, S., Lewis, G., 1999. Support vector machines for optimal classification and spectral unmixing. Ecological Modelling 120 (2-3), 167–179.

Corani, G., Zaffalon, M., 2007a. JNCC2: the Java implementation of the Naive Credal Classifier 2. Tech. Rep. 09-07, Idsia.

Corani, G., Zaffalon, M., 2007b. Naive Credal Classifier 2: a robust approach to classification for small and incomplete data sets. Tech. Rep. 08-07, Idsia.

Duda, R. O., Hart, P. E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Garzon, M., Blazek, R., Neteler, M., Sanchez de Dios, R., Ollero, H., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris l. in the Iberian peninsula. Ecological Modelling 197 (3-4), 383–393.

Nunez, H., Sanchez-Marre, M., Cortes, U., Comas, J., Martinez, M., Rodriguez-Roda, M., Poch, M., 2004. A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. Environmental Modelling & Software 19 (9), 809–819.

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc,US.

Zaffalon, M., 2005. Credible classification for environmental problems. Environmental modelling and software 20 (8), 1003–1012.