Knowledge-Based Systems 66 (2014) 178–189

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Classification of EU countries' progress towards sustainable development based on ordinal regression techniques

M. Pérez-Ortiz [a,*], M. de la Paz-Marín [b], P.A. Gutiérrez [a], C. Hervás-Martínez [a]

[a] Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain
[b] Department of Management and Quantitative Methods, Loyola University, Córdoba, Spain

ARTICLE INFO

Article history:
Received 13 January 2014
Received in revised form 6 April 2014
Accepted 25 April 2014
Available online 9 May 2014

Keywords:
Sustainable development
European Union
Machine learning
Ordinal regression
Ensemble methods

ABSTRACT

Sustainable development (SD) is a major challenge for nations, even more so in the current economic crisis and uncertain environment. Although different indicators, compindices and rankings to measure and monitor SD advances at the macro level exist, the benefits for stakeholders and policy makers are still limited because of the absence of predictive models (in the sense of models able to classify countries according to their SD advances). To cope with this need, this paper presents a first approximation via machine learning techniques. First, we study the SD stage of the 27 European Union Member States using information from the years 2005–2010 and different major indicators that have been related to SD. A hierarchical clustering analysis is conducted, and the patterns are categorised as advanced, followers, moderate and initiated, according to their progress towards SD. The classification problem is addressed from an ordinal regression point of view because of the inherent order among the categories. To do so, a reformulation of the one-versus-all scheme for ordinal regression problems is used, making use of threshold models (Logistic Regression (LR) and Support Vector Machines in this case) and a new trainable decision rule for probability estimation fusion. The empirical results indicate that the constructed model is able to achieve very promising and competitive performance. Thus, it could be used for monitoring the progress towards SD of the different EU countries, in a manner similar to that used for rankings. Finally, the decomposition method based on LR is used for model interpretation purposes, providing valuable information about the most relevant indicators for ranking the end-point variable.

1. Introduction

Sustainable development (SD) is among the most relevant and pressing challenges of the modern age. Since the work of Meadows et al. [25], the interest in this problem has been increasingly growing in the political arena and social consciousness. The underlying idea still remains: human consumption is outstripping what the planet can produce, as we are spending natural resources faster than they can be replenished. In this sense, the academic community, main stakeholders and the political and media debate display special interest in achieving SD as a model of growth for nations and as a primary goal. In times of a deep economic crisis, society and policy-makers focus their attention on economic indicators such as income and employment rates; however, sustainability and social inclusion should also be a priority.

Although there is no consensus in what SD really is (for a discussion of sustainable development definitions see Moldan et al. [27]), a large list of definitions has been published [29,11], and this term is considered a 'contested concept' [18]. However, SD can be said to be known worldwide, thanks to the World Commission on Environment and Development, as development that 'meets the needs of the present without compromising the ability of future generations to meet their own needs' [38].

Nevertheless, what is commonly established is that SD is concerned with ensuring long-term human well-being, which necessarily involves confronting the challenges of limited natural resources and global poverty, having a good standard of living, a long and healthy life, access to education, participation in the social and political life of the community and well-paid work that provides people with the opportunities to achieve their goals, hopes and aspirations [37].

The imperious need for reliable and pertinent indicators, to better monitor and foster SD and to guide this SD process at a

* Corresponding author. Address: Department of Computer Science and Numerical Analysis, University of Córdoba, Rabanales Campus, Albert Einstein Building 3rd Floor, 14071 Córdoba, Spain. Tel.: +34 957 218 349; fax: +34 957 218 630.
E-mail addresses: i82perom@uco.es (M. Pérez-Ortiz), mpaz@uco.es (M. de la Paz-Marín), pagutierrez@uco.es (P.A. Gutiérrez), chervas@uco.es (C. Hervás-Martínez).

http://dx.doi.org/10.1016/j.knosys.2014.04.041
0950-7051/

national level, was recognised early at the time of the Rio Conference and the Agenda 21 [36], followed by the Commission on Sustainable Development work programme on indicators. The most common effect of indicators could be calling attention to an existing problem. However, these indicators yield different scores and rankings depending on the nature and type of assessments. They also report on past performance [3] and they do not predict whether a certain country is heading (or could head) a group in these terms.

A great deal of measurement attempts have been developed over the last two decades at various levels (international organisations, academic and private initiatives) for managing and monitoring progress towards SD [6,10,31,21,30], some of which have focused on households, distribution of wealth, quality of life, social progress and ecological sustainability [4,15,17] but without a consensus on which one is the most determinant at a general level [19].

The fact that governing bodies hold SD as a central strategy reveals the relevance of achieving this development model. In this context, governments need to incorporate SD values and principles in education systems, invest in Research and Development (R&D) to innovate in new technologies (renewal energies, new ways of ecological or organic agriculture, exploitation of scarce natural resources, air contamination reduction, waste treatment, pollution, green buildings, poverty eradication and other such goals) if they wish to progress towards SD and increase human well-being [32]. Currently, different measurement tools, models and methodologies are being used to help stakeholders to analyse the advances of countries in such direction. These tools facilitate the adoption of policies as well as the creation of national systems for a global evaluation.

SD can also be found in the core of EU priorities. According to the Foundational Treaty, the EU institutions work for the achievement of SD in Europe balancing economic growth and price stability, aiming at full employment and social progress and a high level of protection and improvement of the quality of the environment.

The current main instrument of the EU is the long-term Sustainable Development Strategy (EU SDS) [12] and a set of SD indicators developed to evaluate the progress towards SD [13]. These indicators are the main source of information used for the analysis conducted in this paper. There is much in common between well-being and SD indicators approaches and most research on both subjects has concentrated on identifying, developing, and refining criteria and indicators [33].

Among these initiatives, actions to improve and complement the current growth measurements [17] are frequently outlined. This is important, as there are various dimensions that can be interlinked, e.g. education or employment quality can affect health, social relations and status, civic participation, etc. The motivation for proposing an alternative methodology to composite indicators or indices could be that, in the first place, indices summarise too much and provide less information than the description of the characteristics of a cluster or the analysis of models able to predict the class for a new pattern. On the other hand, these indices have been found to be very sensitive to the choices of the index's construction and the selection of the variables to be used.

Because of this need, this paper presents a preliminary study to monitor the progress of the EU toward SD and to validate the different indicators in the literature via different machine learning techniques. The aim of this work can be said to be threefold. The first objective is to conduct a hierarchical clustering analysis to detect behavioural patterns in the EU country-year observations analysed. For this step, 19 Eurostat related official indicators are used, which represent major scenarios and are reported as informative for the progress towards SD in the EU SDS. From this first analysis, we obtained a hierarchical tree of the country-year observations, whose nodes are fused by an expert (examining their characteristics) obtaining four differentiated clusters. Moreover, an ordering for the different resultant clusters is also proposed according to their overall SD performance. In order to obtain a model for deciding the cluster of new countries and to measure and monitor their SD advances, the second stage of the proposed methodology applies ordinal regression to model the categories obtained from the clustering analysis. Given the ordinal nature of the classification problem, the next objective of the paper is to assess the learning problem from an ordinal regression perspective (to benefit from this order information). To do so, we reformulate the decision rule of a recently proposed ordinal ensemble algorithm [28] to obtain more robust results and accurately predict the progress towards SD. The last objective is related to the interpretation of one of the ensemble models obtained in such a way that we can provide valuable information about the most relevant variables in relation to SD progress.

Although the application of a classification algorithm after a clustering process could seem to be trivial, there are three reasons why, in our case, the derived models contribute interesting information: (1) the fact that we are considering expert knowledge to fuse some of the clusters derived from the hierarchical clustering, (2) the ordering given to the resulting groups together with the application of ordinal regression, which provide experts with a ranking tool able to classify new evaluated countries in the groups discovered and to rank them (even when they belong to the same group) and (3) the extraction of information about the most important indicators to characterise and differentiate sustainable developers countries from non-sustainable developers countries (rather than just validating the performance of machine learning algorithms). Note that, for this interpretability analysis, the clustering algorithm itself is not usually very helpful.

Ordinal regression (also known as ordinal classification) problems arise in fields as information retrieval, preference learning, economy, and sociology and nowadays it is considered as an emerging field in the areas of machine learning and pattern recognition. Ordinal regression could be said to be a relatively new learning paradigm which shares properties of classification and regression. Formally, for this problem, $\mathcal{Y}$ (the labelling space) is a finite set, but there exists some ordering among its elements. In contrast to regression, $\mathcal{Y}$ is a non-metric space, thus distances among categories are unknown. Besides, the standard zero-one loss function does not reflect the ordering of $\mathcal{Y}$.

A great number of statistical methods for categorical data treat all response variables as nominal, in such a way that the results are invariant to order permutations on those variables. However, there are many advantages in treating an ordered categorical variable as ordinal rather than nominal [1,22], a statement applicable to classification problems. In this vein, several approaches to tackle ordinal regression have been proposed in the domain of machine learning over the years, since the first methodology (the Proportional Odds Model or POM) dating back to 1980 [24]. Indeed, the most popular approach in this paradigm is the use of threshold models, which are based on the assumption that an underlying real-valued outcomes exist although they are unobservable.

For the learning process, a recently proposed ordinal ensemble methodology [28] based on threshold models is used, and a new trainable decision rule is developed to simplify and make it more robust than the initial proposal. The ensemble methodology is based on decomposing ordinal regression problems into simpler classification tasks, where the order information is explicitly included. The main hypothesis is that the performance of any ordinal algorithm could be improved by simplifying the original classification problem and formulating multiple order hypotheses (hypotheses that will be combined in a final decision function). Therefore, the decision function choice is a crucial step. However,