ARTICLE IN PRESS

JAL:457

Journal of Applied Logic ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Analysis of meteorological conditions in Spain by means of clustering techniques

Ángel Arroyo [a,*], Álvaro Herrero [a], Verónica Tricio [b], Emilio Corchado [c]

[a] *Department of Civil Engineering, University of Burgos, Burgos, Spain*
[b] *Department of Physics, University of Burgos, Burgos, Spain*
[c] *Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
k-means
SOM k-means
k-medoids
Agglomerative hierarchical clustering
Cluster based on Gaussian Mixture Models
Clustering evaluation techniques
Meteorology

ABSTRACT

A comprehensive analysis of clustering techniques is presented in this paper through their application to data on meteorological conditions. Six partitional and hierarchical clustering techniques (k-means, k-medoids, SOM k-means, Agglomerative Hierarchical Clustering, and Clustering based on Gaussian Mixture Models) with different distance criteria, together with some clustering evaluation measures (Calinski–Harabasz, Davies–Bouldin, Gap and Silhouette criterion clustering evaluation object), present various analyses of the main climatic zones in Spain. Real-life data sets, recorded by AEMET (Spanish Meteorological Agency) at four of its weather stations, are analyzed in order to characterize the actual weather conditions at each location. The clustering techniques process the data on some of the main daily meteorological variables collected at these stations over six years between 2004 and 2010.

1. Introduction

Meteorology and climatology are different fields of study, although they may often be confused. Meteorology is the scientific study of atmospheric phenomena, physical processes in the atmosphere and atmospheric effects on the weather. Meteorologists then produce weather forecasts that predict short changes in the weather. In contrast, climatology is the study of atmospheric changes that define average climates and their long-term changes, due to both natural and anthropogenic variations in the climate. Climatological studies therefore share certain meteorological parameters, although climatology predicts long-term weather patterns through climatic models rather than making short-term forecasts. The present study focuses on the clustering analysis of meteorological data from four locations in Spain over a five-year period.

* Corresponding author.
E-mail addresses: aarroyop@ubu.es (Á. Arroyo), ahcosio@ubu.es (Á. Herrero), vtricio@ubu.es (V. Tricio), escorchado@usal.es (E. Corchado).

http://dx.doi.org/10.1016/j.jal.2016.11.026
1570-8683/

Clustering is a useful technique [29] in the study of meteorological phenomena and correct selection of the right clustering algorithm, a requisite for successful experiments. Clustering can be defined as the unsupervised classification of patterns into groups [14]. Hence, clustering (or grouping) techniques will divide a given dataset into groups of similar objects, according to various "similarity" measures. These sets of techniques have previously been applied to meteorological data. In [25], 24-hour mass air trajectories were analyzed at a location in Spain over a three-year period. Clustering techniques with spherical trigonometry were applied, together with the kernel regression method, for their calculation. A multivariate data cube was investigated in [32], to establish whether climate and vegetation classes coincided. To do so, unsupervised clustering techniques were applied and differences between clustering of climate variables versus vegetation variables were studied. In [11], two clustering techniques and a neural network were applied, in an analysis of air quality in Greece and the impact of weather circulation patterns on urban air quality over a period of five-years. Principal Components Analysis (PCA) and Cluster Analysis (CA), were applied in [26] over a 3-year period to analyze the mass concentrations of Sulfur Dioxide (SO_2) and Particulate Matter (PM10) in Oporto. Finally, a clustering method for the study of multidimensional non-stationary meteorological time series was presented in [15] and the results were compared with standard fuzzy clustering techniques for a dataset with temperatures in Europe over forty years. In this study, unlike previous works, cluster evaluation measures, together with partitional and hierarchical clustering techniques, based on different distance measures, are employed to categorize different climate zones in Spain.

So, with these promising techniques, we can analyze the study of both the similarities of the main regional climates and their differences. The climate in Spain is highly variable, mainly due to its position in southern Europe, diverse relief, and extensive coastline. The Iberian Peninsula is in a temperate zone where currents of warm air and cold air merge to create its unique meteorological conditions. This great variability of climatic zones means that Spain is a European country of special interest for a meteorological study of the sort proposed in this study. Its various climatic subtypes [8] are reflected in the data gathered from four points: a Mediterranean island; an interior location on the meseta of the Iberian Peninsula; a city on the southern coastline; and, a city to the north-west of the Iberian Peninsula. The network of weather stations for meteorological data acquisition are constantly recording continuous data streams that are publicly accessible for research and analysis [23]. Described in detail in Section 3, these stations represent points within each of the four main Spanish climatic zones (continental, Atlantic, dry Mediterranean and typical Mediterranean).

Unlike the time window in a previous work [3] by the authors, a wider time window is used for data analysis in this study, running between 2004 and 2010. Additionally, a larger and more comprehensive set of techniques analyses the meteorological the extensive time series of data from the four different climatic zones (see Section 3). Firstly, four cluster evaluation measures yielded an accurate estimation of the recommended number of clusters for the dataset. Secondly, various clustering techniques applied to the original data set allowed us to assign the best possible data clustering technique. Four relevant partitional [2] techniques, one hierarchical [21] technique, and four cluster evaluation measures [17] were applied, combined with the most widely-used distance measures. The results were analyzed in two ways: through a study of the meteorology at the four selected locations and through clustering technique comparisons to establish the advantages of each method.

The rest of this paper is organized as follows. Section 2 presents the clustering techniques, distance criteria, and the cluster evaluation measures applied. Section 3 describes the real-life case study and, Section 4, the experimental results. Finally, Section 5 sets out the main conclusions and future lines of work.

2. Clustering techniques

This study reports the performance of several clustering techniques analyzing time series of data on meteorological conditions (described in Section 3), studying the climatology of different locations. Several