Journal of Atmospheric and Solar-Terrestrial Physics 120 (2014) 121–131

Contents lists available at ScienceDirect

Journal of Atmospheric and Solar-Terrestrial Physics

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

Clustering of atmospheric data by the deterministic annealing

Alexander Ruzmaikin, Alexandre Guillaume

Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109, USA

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form
9 September 2014
Accepted 10 September 2014
Available online 18 September 2014

Keywords:
Satellite data
Clustering
Temperature
Water vapor
Deep convective clouds

ABSTRACT

The Deterministic Annealing (DA) clustering method, which determines the cluster centers, their sizes, and probability with which data are associated with each cluster, is tested using artificial data and applied to atmospheric satellite data. It is also shown how the method can be advantageously used to characterize data outliers. The method is based on the optimization of a cost function that depends both on the averaged distance of data points to cluster centers and the Shannon entropy of the data. The cost function uses two independent parameters in a close analog to the Gibbs' thermodynamics (with the averaged distance similar to the internal energy) allowing a sufficient control of the formation of new clusters as “phase transitions” by changing the clustering parameter similar to the thermodynamical temperature. The satellite data used are a temperature–water vapor data set and the positions of deep convective clouds obtained from the measurements of the Atmospheric InfraRed Sounder (AIRS) on the Aqua satellite. The clustering of these data is demonstrated for the 2D case (at fixed pressure level) and for the 3D case at multiple pressure levels indicating potential applications to investigation of distributions of atmospheric profiles.

Published by Elsevier Ltd.

1. Introduction

Current research in atmospheric and climate sciences relies on the use of data products provided by numerous satellites orbiting the Earth. It is a common practice to arrange the original satellite data into grid boxes of selected size by averaging all measured data points inside a grid box (taking a mean) and providing a standard deviation (see for example http://disc.sci.gsfc.nasa.gov/AIRS/data_holdings/). This practice is based on the assumption that the data distribution within those boxes is normal. This assumption is not always valid, c.f. Perron and Sura (2013), and our direct checks of actual distributions of data in grid boxes show that often the data are distributed in a non-Gaussian manner, *i.e.* that they cannot be accurately characterized by their mean and standard deviation. Adding higher statistical moments (skewness, kurtosis, …) may be useful under the assumption that the data have a unimodal distribution function but not reliable since in some grid boxes the data distribution may be multi-modal. A better approach advocated here is to group grid box data into clusters with characteristics that provide more extended statistical information on measured quantities. The process of agglomerative (and divisive) clustering consists in partitioning a selected data set into several subsets called clusters. In mathematical terms, the problem of partitioning can be cast as optimization of a cost function that characterizes how similar the data in a cluster are to each other compared to the data in other clusters. Clustering methods allow one to evaluate the relative importance of each cluster using simple descriptive statistics. Clustering can also be understood as a way of construction of data histograms with information about the probability distributions underlying their bins and with an additional advantage of applying it to multi-dimensional data.

Cluster analysis is successfully used in atmospheric and climate sciences for analyzing data and model outputs (c.f. Reljin et al., 2002; Steinbach et al., 2003; White et al., 2005; Hoffman et al., 2005). The book by Wilks (2005) gives a description of techniques and a list of applications in atmospheric sciences. The International Satellite Cloud Climatology Project (ISCCP) used a cluster technique to identify cloud and weather regimes (Jakob and Tselioudis, 2003; Rossow et al., 2005). A cluster analysis had been applied to the CloudSat data from the A-train formation of satellites for identifying the type of clouds (Sassen and Wang, 2008). These and most other applications used a popular in climate and atmospheric studies clustering algorithm called K-means (Hartigan, 1975). The K-means cost function uses *one* cluster parameter, the distance measure between a data point and the cluster center. K-means results depend on the selection of the initial centers of clusters (the means) and there is no guarantee that it will converge to a global optimum. A relatively new

E-mail addresses: Alexander.Ruzmaikin@jpl.nasa.gov (A. Ruzmaikin), Alexandre.Guillaume@jpl.nasa.gov (A. Guillaume).

http://dx.doi.org/10.1016/j.jastp.2014.09.009
1364-6826/Published by Elsevier Ltd.

algorithm, the entropy-constrained vector quantization, ECVQ, clustering has been designed for more detailed evaluation of the cluster relative weights or priors (Chou et al., 1989). It does so by using an additional constraint (entropy index) in the cost function. The ECVQ algorithm has been proposed to be used for reduction of the size and complexity of massive climate data (Braverman, 2002; Braverman et al., 2003).

Here as a next step in development of methods of data clustering preserving the information containing in the original data, we consider a more advanced algorithm called Deterministic Annealing (Rose, 1998), which is based on the minimization of the cost function relative to *two* independent parameters and provides probabilities with which data are associated with each cluster. The method has a close and deep analog to the classical Gibbs' thermodynamics based on the use of a minimum two basic variables to define a thermodynamic state, c.f. Landau and Lifshitz (1980).

In Section 2, we briefly describe the Deterministic Annealing (DA) method referring the reader to the original paper by Rose (1998) for technical details. Section 3 illustrates the application of the DA algorithm to artificial data. In Section 4, we present the applications of the DA algorithm to clustering atmospheric data retrieved from the measurements by the Atmospheric Infrared Sounder (AIRS) and to identifying data outliers. Finally, we summarize our results and discuss other potential applications of the algorithm (Section 5).

2. Deterministic annealing clustering

The Deterministic Annealing (DA, Rose, 1998) uses a probabilistic framework for clustering: The input data $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ are assigned to clusters with centers $\mathbf{y} = [y_1, y_2, \ldots, y_K]$ using the conditional probability $p_a = p(\mathbf{y}|\mathbf{x})$ (probability of $\mathbf{y}$ given $\mathbf{x}$), which is called *the association probability*. The algorithm searches for the optimum of the cost function $F = D - \lambda H$, where

$$D = \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y})\, d(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathbf{y}} p_a\, d(\mathbf{x}, \mathbf{y}), \tag{1}$$

$$\begin{aligned} H &= -\sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}, \mathbf{y}) \\ &= -\sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathbf{y}} p_a \log p_a + constant \end{aligned} \tag{2}$$

are the average distance and the Shannon entropy (Shannon, 1948) scaled by a Lagrange multiplier λ. Here $p(\mathbf{x})$, $p(\mathbf{x}, \mathbf{y})$ are the data probability and the joint probability, and $d(\mathbf{x}, \mathbf{y})$ is a distance measure. In this paper, we assume that $d = \| \mathbf{x} - \mathbf{y} \|^2$. The choice of the cost function F is made to restrict randomness in minimization of D relative to the free parameters $\mathbf{y}$, p_a to a level measured by the Shannon entropy. Note that the second lines in Eqs. (1) and (2) show that the cost function depends only on these two free parameters, given the data and data probability distribution.

To show a systematic increase of the information when the parameter λ and the averaged distance decrease, we introduce the rate R that quantifies the information amount in a slightly different fashion than the entropy H does, thereby allowing a clear graphical depiction of the tradeoff (Rose, 1994). Within the context of information theory, R, expressed as a function of D, is "the effective rate at which the source produces information subject to the constraint that the user can tolerate an average distance of D" (Berger, 1971). In the context of the present work and in a loose sense, the rate is a quantity that characterizes the information content of a given solution (set of clusters) with respect to the original data. We calculate R using the following equation:

$$R = \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathbf{y}} p_a \log\left(\frac{p_a}{p_c}\right). \tag{3}$$

where $p_c = p(\mathbf{y})$ is the probability distribution of cluster centers.

Minimizing F with respect to the association probability under the additional constraint that the total contribution of each cluster weight is constant (the so-called mass constrained version of DA) results in the Gibbs distribution:

$$p_a = \frac{p_c e^{-d(\mathbf{x},\mathbf{y})/\lambda}}{Z_x}, \tag{4}$$

where $Z_x = \sum_{\mathbf{y}} p(\mathbf{y}) e^{-d(\mathbf{x},\mathbf{y})/\lambda}$. Minimization of F with respect to the cluster locations $\mathbf{y}$ leads to the condition:

$$\sum_{\mathbf{x}} p_a \frac{d}{dy} d(\mathbf{x}, \mathbf{y}) = 0, \tag{5}$$

where p_a is the Gibbs distribution (4).

The DA notation has a close analog in thermodynamics with F similar to the Helmholtz free energy. The averaged distance D plays the role of the internal energy and the Lagrange multiplier λ plays the role of the temperature in thermodynamics. The minimum of the free energy determines the distribution of a system at thermal equilibrium, *i.e.* the Gibbs distribution (4). The procedure of annealing consists of maintaining the system at thermal equilibrium while carefully lowering the temperature. This physical analogy helps to understand the work of the DA algorithm and to control the size of the clustering model by observing the "phase transitions" (formation of new clusters) at some critical values of λ, see below.

The algorithm is called "deterministic" because the cost function is directly minimized in contrast to some other clustering algorithms, such as K-means, that rely on stochastic simulations. The algorithm entails a gradual decrease of λ to find a minimum of F at each λ, *i.e.* effectively "annealing" the system. Technically, for each value of λ the association probability (4) is calculated for a fixed set of clusters. Subsequently, the clusters location is updated using Eq. (5) at each value of the association probability.

It can be shown that high λs imply that the global minimum of F is found (Rose, 1998). When $\lambda \to \infty$, the probability (4) is uniform and the condition (5) leads to a single cluster, the sample mean of the data. The number of clusters increases as λ decreases undergoing "phase transitions" (formation of clusters). When $\lambda \to 0$ each data point becomes a one-point cluster. A simple method applied to the minimum F allows us to determine a "critical temperature" λ_c at which a phase transition, *i.e.* appearance of a new cluster, occurs (Rose, 1998). Specifically, at each new λ, the condition for phase transition $\lambda \le \lambda_c$ is checked for each cluster. The critical λ_c is calculated as twice the value of the largest eigenvalue of the covariance matrix of the posterior probability $p(\mathbf{x}|\mathbf{y})$:

$$C_{\mathbf{x}|\mathbf{y}} = \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^{tr}, \tag{6}$$

where the symbol 'tr' indicates the transpose matrix. If the condition is found true for one cluster, another centroid is added. The formation of a new cluster takes place when the value $\lambda_c = 2\lambda_{max}$ is reached, with λ_{max} being the largest eigenvalues of the covariance matrix.

3. How the DA algorithm works

Let us first show how the DA algorithm works using our Matlab implementation of the algorithm. In this implementation we enter