Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 6 (2011) 231–236

Procedia
Computer Science

Complex Adaptive Systems, Volume 1
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2011- Chicago, IL

Identification of severe weather outbreaks using kernel principal component analysis

Andrew E. Mercer[a*], Michael B. Richman[b], Lance M. Leslie[b]

[a]Northern Gulf Institute, Mississippi State University, 108 Hilbun Hall Mississippi State, MS 39760 USA
[b]University of Oklahoma School of Meteorology, 120 David L. Boren Blvd., Norman, OK 73072 USA

Abstract

A new adaptive approach to severe weather outbreak compositing and discrimination is described for datasets of known non-tornadic and tornado outbreaks. Kernel principal component analysis (KPCA) is used to reduce the dimensionality of the dataset and provide input for cluster analysis (CA) of the outbreaks to discern meteorological characteristics unique to each outbreak type. Results are compared to traditional principal component analysis (PCA). The KPCA methodology and CA assigned outbreaks to different composite (maps that have a close correspondence) sets than did PCA and CA. The clusters associated with each method were used as training for a support vector machine classification scheme. An independent subset of the outbreak dataset was retained for cross-validation classification of outbreak type. Significant differences in the two composite methods are observed, and a support vector machine classification scheme demonstrates compelling effectiveness in distinguishing outbreak types based on the resulting composites.

Keywords: Principal Component Analysis; Kernel Methods; Support Vector Machines; Severe Weather Outbreaks

1. Introduction

Outbreaks of severe weather, which involve multiple potentially destructive storms in a single system, adversely affect the central and eastern United States annually, causing millions of dollars in damage and numerous fatalities. Owing to their destruction, these events are catalogued in a research data base. One of the attributes of this dataset is the type(s) of severe weather produced (e.g., tornadic or nontornadic). Doswell et al. (2006) define a tornado outbreak (TO) as a storm system with at ≥ 6 tornadoes encompassing a space scale >1000 km in diameter (known as synoptic scale). Their methodology distinguishes TOs from primarily nontornadic outbreaks (PNTOs) and ranks each outbreak based on the relative severity, using variables such as the number of significant hail reports, the number of significant wind reports (> 33 ms^{-1}), the number of significant tornadoes, and the number of fatalities.

1877–0509 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.procs.2011.08.043

The ranking is used to create TO and PNTO types or classes. Translating this ranking scheme into a physical understanding of atmospheric conditions that give rise to these high impact events requires understanding the large-scale physical processes that govern each type of outbreak with a lead time useful for decision makers. Previous research on large-scale processes is limited to case studies of major TOs (e.g. Roebber et al. 2002), or averages of a series of maps (Schaefer and Doswell 1984). No study has investigated synoptic-scale conditions unique to TOs and PNTOs, a prerequisite for successful forecasting of severe event type. Mercer et al. (2009) formulated a support vector machine (SVM - Cristianini and Shawe-Taylor 2000) that distinguishes TOs from PNTOs using output from a synoptically forced numerical weather prediction model. However, that study did not identify synoptic scale precursors to outbreak type, rather focusing on mesoscale (on the order of 100 km) processes in numerical simulation output useful for discriminating outbreak types. Identifying synoptic patterns spanning each outbreak type requires synoptic composites for a 3-D atmospheric domain, centered on particular outbreak events.

This study formulates synoptic composites of TOs and PNTOs using principal component analysis (PCA). However, one major limitation of PCA is the assumption of linearity. This is particularly problematic when considering highly nonlinear atmospheric data; therefore, kernel principal component analysis (KPCA - Schölkopf et al. 1998) is applied to the outbreak data. Richman and Adrianto (2010) demonstrated the utility of KPCA in formulating atmospheric height patterns over Europe. Their methodology is applied here to the TO and PNTO data. The main goal of this work is to appraise the differences between PCA and KPCA formulated composites of TOs and PNTOs. The outbreak discrimination capabilities of the composites are assessed using a SVM. Section 2 outlines the data and methodology. Results are show in Section 3 and summarized in Section 4.

2. Data and Methodology

2.1 Data

As the goal of this study is to determine the meteorological signals associated with different types of severe weather outbreaks, datasets of outbreaks and a synoptic-scale meteorological dataset are required. The 50 highest ranked TOs and PNTOs from 1970 – 2005, using to the Doswell et al. (2006) methodology, were obtained (a list is available in Shafer et al. 2009). These 50 events were deemed to be the most prototypical and distinct between the two classes. Failure to observe differences in the meteorological fields associated with these two sets of events would suggest further investigations are unwarranted.

The synoptic-scale dataset used to formulate the composites is a collection of surface and atmospheric information collected from 1948 to the present and processed through a specific analysis model to produce a reanalysis (Kalnay et al. 1996). These data are available on a 2.5° latitude-longitude grid with 17 vertical levels, and a surface level. Kalnay et al. (1996) provide a reliability grade for different variables in the reanalysis, based on their use of observations in lieu of model and/or climatology data. An A grade is the most reliable and D the least. To portray the 3-D atmospheric state, five reanalysis quantities were used that describe atmospheric energy content (temperature – reliability of A), atmospheric moisture (relative humidity - B), and atmospheric motion (u and v wind components) and pressure - A. At upper levels, the height of a constant pressure surface is used instead of vertical pressure values and is quantified as geopotential height (A).

2.2 Composite Methodology

The goal of compositing methodology is to isolate highly similar patterns in a data matrix, **X.** For meteorological applications, the matrix **X** consists of a parameter dimension (e.g. temperature, relative humidity), a spatial dimension (gridpoint observations), and a temporal dimension (individual case days). The matrix **X** is a 3D cube, whereas the analyses are 2D, so the data cube must be flattened into a 2D matrix. In this study, the same weather variables are used for each event; therefore, rows of the matrix index gridpoint observations and columns index individual case days. Because reanalysis data are provided on a latitude-longitude grid, longitude convergence occurs moving away from the equator that will inflate artificially, as a function of latitude, the similarity values used for both PCA and KPCA. To avoid convergence, the data are mapped to a new set of equally spaced gridpoints, known as a Fibonacci grid (Swinbank and Purser 2006), with only a 1% interpolation root mean square error introduced by this projection. After the data are interpolated onto the Fibonacci grid, defining **X,** a similarity matrix is formed (for PCA, a correlation or covariance matrix; for KPCA, a kernel matrix). This similarity matrix describes