Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 48–57

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Using self-organizing maps for unsupervised analysis of radar data for nowcasting purposes

Gabriela Czibula[a], Andrei Mihai[a,*], Eugen Mihuleţ[b], Daniel Teodorovici[a]

[a]Department of Computer Science, Babeş-Bolyai University
1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania
[b]National Meteorological Administration
Şos. Bucureşti-Ploieşti nr.97, 013686, Bucharest, Romania

Abstract

Predicting weather, and particularly severe weather, is an important challenge both for meteorological and machine learning researchers. The complexity and difficulty of the problem is mainly due to the chaotic character of the atmosphere and the implicit large set of meteorological information (radar, satellite or ground meteorological observations) which have to be analyzed by meteorologists. Thus, understanding the relationships between various meteorological parameters extracted from radar observations may be useful for providing additional comprehension about severe weather development and would help to identify situations when severe weather can occur. *Self-organizing maps* are being explored as an unsupervised classification model for detecting patterns in radar data which are relevant in predicting short-term weather changes. Experiments are performed on real radar data provided by the Romanian National Meteorological Administration. With the main goal of analyzing how the values for the weather radar products are evolving between consecutive radar scans, we empirically show that in general there is a slow change in the values over time, except for the situations when certain severe phenomena occur. The study conducted in this paper is aimed to provide a better insight regarding how the values of weather radar products are evolving in time both in calm and severe weather conditions, with the broader goal of using these findings for weather nowcasting.

Keywords: Nowcasting; Unsupervised learning; Self-organizing map
2000 MSC: 37N10; 68T05; 62H30

1. Introduction

As stated by the World Meteorological Organization (WMO) [24] weather, and particularly severe weather, causes many natural disasters and is responsible for lots of damages and loss of life. Since the number and intensity of severe

[*] Corresponding author. Tel.: +4-264-405-327; fax: +4-264-591-906.

[*] Corresponding author. Tel.: +4-264-405-327; fax: +4-264-591-906.
E-mail address: mihai.andrei@cs.ubbcluj.ro

10.1016/j.procs.2019.09.159

weather events is increasing in various regions of the world, the problem of forecasting such phenomena and issuing early warnings is nowadays one of the most popular topics in meteorology. In particular, *weather nowcasting* which is the analysis and forecast for the next 0 to 6 hours is of major interest within the meteorological research. The problem of issuing a nowcasting warning can be very difficult for meteorologists, since there is often an extremely large set of meteorological data (available in the form of radar, satellite or ground meteorological observations) which has to be analyzed in a very short period of time and the constraints imposed on a desirable solution are complex and not exactly known. Therefore, *machine learning* (ML) based methods (particularly the *supervised* ones) are necessary for obtaining effective solutions for the nowcasting problem. In addition, the *unsupervised* learning methods are useful for extracting accurate and meaningful patterns from the large amount of weather related data and to improve decision-making for high-impact weather.

With the broader goal of contributing to the enhancement of national nowcasting warning systems, we are performing in this paper an unsupervised analysis of the meteorological data which is recorded, at the national level, in the form of radar observations and used for weather nowcasting. Investigating unsupervised learning models for providing an insight about the data organization and structure represents a natural preliminary step before the application of a predictive supervised learning model. With this aim, as a proof of concept, we are assessing the usefulness of *self-organizing maps* (SOMs) to unsupervisedly uncover the underlying structure in radar data, for analyzing how the values for the main weather radar products are evolving between consecutive radar scans and for studying the relevance of the radar products in predicting short-term weather changes. Through several experiments performed on real radar data provided by the Romanian National Meteorological Administration (NMA) and collected from the Central Transylvania region we aim to obtain an empirical evidence that the radar meteorological products' values are generally smoothly changing in time in normal weather conditions, excepting situations when certain severe phenomena occur. In addition, we expect that SOMs are able to distinguish severe weather conditions from radar data. If this stands, we can advance our research towards applying deep learning models [5] for predicting the near future values for the radar products based on their historical values. The use of the self-organizing maps is not new for the analysis of weather and the evaluation of other meteorological activities. However, to the best of our knowledge SOMs have not been investigated in a general nowcasting context, with the aim of unsupervisedly detecting dependencies in radar data that would be helpful for the short-term prediction of weather.

To summarize, in this paper we aim to find answers to the following research questions:

RQ1 What is the potential of self-organizing maps to unsupervisedly detect some patterns in the way the meteorological products are modifying for short time periods (e.g. 6 minutes), particularly in situations when certain severe phenomena occurred?

RQ2 What are the meteorological products which are the most relevant for detecting severe weather conditions?

RQ3 Are SOMs able to capture meteorological relevant patterns? More specifically, are our computational findings obtained by answering RQ1 and RQ2 correlated with the meteorological perspective?

We expect that the answers to the previously stated research questions will provide us additional insights about the best way for modelling the nowcasting problem as a supervised learning problem.

The rest of the paper is structured as follows. Section 2 briefly presents the nowcasting problem and the unsupervised SOM model used in the study. A literature review on using unsupervised learning methods for analyzing meteorological data is presented in Section 3. Section 4 introduces a way of modelling the radar data and our methodology regarding the use of SOMs as descriptive models for uncovering the underlying structure of the meteorological information collected by radar. Section 5 presents the experiments performed following the methodology introduced in Section 4, including the description of the data set used, a statistical analysis of the data sets and the experimental setup. Section 6 provides the obtained results and their analysis, both from a computational and a meteorological perspective. The conclusions of our paper and directions to further continue our research are given in Section 7.

2. Background

This section starts by introducing the importance of the nowcasting problem. Then, a brief background on *self-organizing maps* is presented.

2.1. Weather nowcasting

According to a recent joint article from Nordic and Baltic countries [19], climate change including extreme rain phenomena is expected. In consequence, there is an increasing need for accurate and early warning of severe weather events. As the number and intensity of severe meteorological phenomena increases, predicting them in due time to avoid disasters becomes highly demanding for meteorologists. The division of weather prediction dealing with weather analysis and forecast for the next 0 to 6 hours is called *nowcasting* and plays an increasing role in crisis management and risk prevention. Meteorological institutes hold a large set of historical meteorological data, such as radar measurements, satellite data and ground observations, all accessible to be processed. In addition, weather-focused satellites currently orbit the Earth, tracking data about cloud patterns, winds, temperature, while radars and ground stations are constantly gathering real-time data. Thus, a lot of data from various sources may be used by *machine learning* (ML) based algorithms for improving the performance of weather-prediction techniques. Given the amount of data available to be processed, as well as the need for improved nowcasting techniques, there is a high opportunity to analyze and accurately predict weather phenomena.

The problem of issuing a nowcasting warning is a difficult task for meteorologists, mainly because of the extremely large set of data which has to be analyzed in a short period of time. Therefore, ML based methods are useful for offering effective solutions for nowcasting by learning relevant patterns from the large amount of weather data and thus improving decision-making for high-impact weather. Most of the existing operational and semi-operational methods for nowcasting are using the extrapolation of radar data and algorithms mainly based on cell tracking. As previously described, existing nowcasting techniques use various data sources which may be relevant for accurate nowcasting, such as: meteorological data (radar, satellite, meteorological observations) and geographical data (elevation, exposure, vegetation, hydrological features, anthropic features). The current study uses radar data provided by the WSR-98D weather radar [16]. The WSR-98D is a type of doppler radar capable of detecting water droplets in the atmosphere, retrieving data on particles location, size and motion, this data being delivered to regional and national meteorological centers in a particular data format, i.e. NEXRAD Level III . About every 6 minutes data is collected on a complete set of about 30 base and derived products, gathered over 7 different elevations. The base products are particle *reflectivity* (R), providing information on particle size and type, and *particle velocity* (V), containing information on particle motion. Both products are available for several elevation angles of the radar antenna, and for each time step a set of seven data products, R01-R07 and V01-V07, is delivered, each of them corresponding to a certain tilt of the antenna. Among the derived products, of particular interest for the study is VIL (vertically integrated liquid), an estimation of the total mass of precipitation above a certain unit of area. The data in the NEXRAD Level III files is stored in a gridded format, each point of the grid corresponding to a geographical location and containing the value of a certain product at the respective time frame. In the data grid provided by the WSR-98D radar, the OX axis contains the longitude values, while the OY axis contains the latitude values.

2.2. Self-organizing maps

A *self-organizing map* (SOM) [18] is an unsupervised learning model, a type of artificial neural network from the category of *competitive learning* networks. A SOM contains two layers: the input layer and the output layer. Each neuron from the input layer of a SOM is connected to each neuron from the output layer and each connection has an associated weight. The output layer, also called a *map* [4], represents a low-dimensional (usually two-dimensional) representation of the training samples. The map preserves the topological relationships in the input space, in other words similar input instances are grouped on map neurons (units) that are close together [9]. Usually, a SOM is trained using the Kohonen algorithm [18]. While a SOM is considered a tool for visualizing high dimensional data, it is also very effective for clustering problems, as a trained self-organizing map provides clusters of similar data items [10]. Thus SOMs are appropriate for data-mining tasks involving clustering or classification [10].

The U-Matrix method [7] is usually used for visualizing a trained SOM. Each node from the resulting map has an associated U-Matrix value computed as the average of the Euclidean distances between the node and its closest neighbors (usually four or eight). Considering the U-matrix values as heights in a landscape, the U-Matrix may be interpreted as follows [7]: low values on the U-matrix encode similar input instances which can be grouped together to represent data clusters, while high places represent separation boundaries between clusters of instances.

3. Literature review

While there is no study in the literature similar to the one performed in this paper, there are previous works in which unsupervised learning methods have been used in a nowcasting context. SOMs were already applied in meteorological tasks for data analysis or data preprocessing purposes.

Vilibic et al. [23] present a method for nowcasting and short-term forecasting (up to 72 hours) of ocean currents based on the forecasts of winds. The proposed method consists of training the SOM using data created from joining surface wind pattern data to surface ocean currents data. The method was tested on three real data sets from the Adriatic Sea and compared to the existing ocean physics based system. The SOM-based method proposed in [23] was shown to be slightly better than the existing method, providing an 8.7% lower RMSE score. Lepioufle presents in [11] an extensive feasibility study centered on quantitative precipitation estimation. In this study, Lepioufle mentions SOMs as being useful for data preprocessing and binary classification. The data is gathered from a meteorological radar, for prediction, and from hourly automatic gauges, for estimation correction.

Weisberg and Liu [13] describe several applications of SOMs in meteorology analyzing data such as rainfall, evaporation, air temperature, humidity and sea level pressure. several applications in oceanography are also presented, analyzing data such chlorophyll, geochemical or sea surface temperature data.

Tambouratzis and Tambouratzis [20] present a method of classifying areas in Greece into meteorological profiles using SOMs and statistical analysis. The data is gathered from 130 weather stations and represents weather data spanning 43 years, having 28 different parameters. Each parameter was averaged over the 43 years for every weather station. A SOM was used in order to cluster the weather stations into groups of stations represented by similar meteorological characteristics. They used a map size of 2x5, resulting in 10 sets of meteorological characteristics, then Ward's statistical clustering was used to combine similar sets resulting in a total of 4 meteorological profiles. The authors show that the method is consistent in classifying stations into similar meteorological profiles as well as successfully classifying novel stations, not used during training. The same authors then extend this work in [21]. They performed better data preprocessing: eliminating 2 of the 130 weather stations because of missing data concerns; performing parameter selection in order to eliminate highly correlated parameters, while preserving the original distance relations of the data set, resulting in a selection of 20 parameters out of 28; data normalization, comparing range and variance normalization. The authors also analyzed the topology preservation capability of the SOM for different map sizes. The best results were obtained using a map size of 3x5 which resulted in 3 different geographical profiles.

Kalinic et al. [6] applied SOMs for obtaining characteristic surface wind patterns using two meteorological models operational in the northern Adriatic, Aladin/HR and WRF-ARW. Experiments were conducted on a data set consisting of hourly values for the surface wind collected from 1 February to 30 October 2008. The results revealed that the unsupervised SOM model was able to distinguish between the main types of the northern Adriatic winds: bura, jugo and maestral-tramontana [6]. Lin et al. [12] introduced a hybrid learning model for forecasting reservoir inflow during typhoon periods. The proposed model used a support vector regressor whose training instances were previously clustered using an unsupervised SOM model. Rainfall data collected from 1988 to 2008 on the Feitsui Reservoir watershed from northern Taiwan was used in the experimental evaluation of the SOM-SVM model. The experimental evaluation revealed an improved forecast of hourly inflow using the hybrid SOM-SVM model.

A recent study performed by Ohba et al. [17] applied an ensemble of SOMs for probabilistic prediction of local precipitations in Japan. SOMs were used to downscaling medium-range ensemble forecasts by detecting relationships between local precipitation data and certain atmospheric patterns. Experiments conducted on atmospheric data from Japan highlighted that SOMs provided a significantly improves predictive performance of the ensemble forecasts [17].

4. Methodology

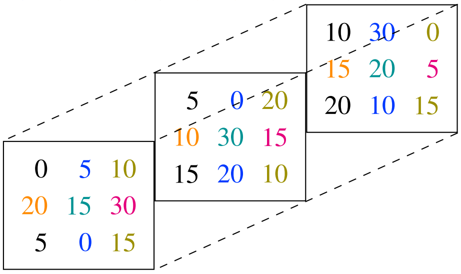
We further detail the methodology used for investigating the effectiveness of SOMs in analyzing radar data for unsupervisedly detecting severe weather events.

As previously shown in Section 3, the exported raw data collected through the radar scans during one day (24h) on a certain geographic region is provided as a sequence $\mathcal{S}$ of m x n dimensional matrices. A matrix M from $\mathcal{S}$ corresponds to a certain time stamp t and a meteorological product p (e.g. R01). We denote by np the number of meteorological products provided by the radar. An element a_{ij} from M represents the value for the product p on the cell (i,j) from the m x n grid corresponding to the analyzed region. The sequence $\mathcal{S}$ of matrices may be visualized as a 3D data grid. Let us denote by $t_1^d, t_2^d, \ldots, t_k^d$ the time stamps for which radar data is recorded on a certain day d, where by t_1^d we express the time when the radar started to collect data. Assuming that radar data is collected every 6 minutes during one day, the number *k* of time stamps is equal to 240. For each time moment *t*, a sequence of matrices (3D data grid) is available, containing the values for various radar products at time *t*.

4.1. The proposed data model

We propose in the following a data model which will be further used in our experiments. The idea is to assign, at each time stamp, a vectorial representation to each 3D data grid provided by the radar. In this model, for a day *d*, a time stamp t_i^d $(1 \leq i \leq k)$ and a set *Prod* of meteorological products, a data parallelepiped $P_{t_i^d}(m, n, Prod) = (p_{xyz})_{\substack{x=\overline{1,m}\\ y=\overline{1,n}\\ z=\overline{1,|Prod|}}}$ is constructed. In this parallelepiped, *OX* and *OY* axes represent the rows and columns from the radar data grid, and the depth axis *OZ* represents the meteorological products. For obtaining the vectorial representation for the data parallelepiped $P_{t_i^d}(m, n, Prod)$, it is linearized as follows. First, the 3D *m* x *n* x *np* parallelepiped is converted into a 2D *m* x (*n*·*np*) matrix as follows. The sequence of *np* components (from the *OZ* dimension) of each element (*x*,*y*) from the *x*-th row $(1 \leq x \leq m)$ and *y*-th column $(1 \leq y \leq n)$ of the parallelepiped $P_{t_i^d}(m, n, Prod)$ is incorporated in the *x*-th row of the 2D matrix. The second step is to liniarize the 2D *m* x (*n*·*np*) matrix into an (*m*·*n*·*np*)-dimensional vector obtained by concatenating the rows from the matrix (from the first row until the *m*-th row). Consequently, after the linearization process, the parallelepiped $P_{t_i^d}(m, n, Prod)$ is transformed into a vector $V_{t_i^d}(m, n, Prod) = (v_1, v_2, \ldots, v_{m \cdot n \cdot np})$, such that $v_{(x-1)\cdot n \cdot np + (y-1)\cdot np + z} = p_{xyz}$ $\forall 1 \leq x \leq m, 1 \leq y \leq n, 1 \leq z \leq np$.

For exemplifying the proposed data model, let us consider that the dimensions of the grid are $m = 3, n = 3$ and the set *Prod* of meteorological products is $Prod = \{R01, R02, R03\}$. Figure 1 depicts a sample data parallelepiped $P_t(m, n, Prod)$ at a certain time stamp *t*. The matrix in front from Figure 1 contains values for R01 for each cell from the 2x3 data grid, the middle matrix represents values for R02 and the matrix from behind consists of values for R03. As previously shown, the 3D data grid (Figure 1) is first converted to a 2D matrix (Figure 2). Then, the second step is to liniarize the resulting 2D matrix. The vector $V_t(m, n, Prod)$ built for the data parallelepiped from Figure 1 $P_t(m, n, Prod)$ using the proposed data model is (0, 5, 10, 5, 0, 30, 10, 20, 0, 20, 10, 15, 15, 30, 20, 30, 15, 5, 5, 15, 20, 0, 20, 10, 15, 10, 15).



$$\begin{matrix} 10 & 30 & 0 \\ 15 & 20 & 5 \\ 20 & 10 & 15 \end{matrix}$$

$$\begin{matrix} 5 & 0 & 20 \\ 10 & 30 & 15 \\ 15 & 20 & 10 \end{matrix}$$

$$\begin{matrix} 0 & 5 & 10 \\ 20 & 15 & 30 \\ 5 & 0 & 15 \end{matrix}$$

Fig. 1: A sample 3D data parallelepiped $P_t(m, n, Prod)$

$$\begin{matrix} 0 & 5 & 10 & 5 & 0 & 30 & 10 & 20 & 0 \\ 20 & 10 & 15 & 15 & 30 & 20 & 30 & 15 & 5 \\ 5 & 15 & 20 & 0 & 20 & 10 & 15 & 10 & 15 \end{matrix}$$

Fig. 2: The 2D matrix corresponding to the 3D data grid from Figure 1

4.2. Experiment

With the goal of answering research questions RQ1 and RQ2 formulated in Section 1, several case studies are considered using historical data provided by NMA and the data model previously introduced. The experiments are aimed to analyze the extent to which SOMs are able to unsupervisedly detect severe meteorological phenomena.

Two data sets *D*1 and *D*2 are constructed for representing the radar data collected during the time stamps $t_1, t_2, \ldots, t_k$ on a day *d* using the data model introduced in Section 4.1. Thus, the data sets consist of *k* instances, each instance representing the vectorial representation *V* (Section 4.1) of the grid at certain time moment *t*. The difference between *D*1 and *D*2 is given by the set of meteorological products used for representing the instances. In *D*1 the entire set of meteorological products provided by the radar (i.e. 24) is used, while *D*2 employs only 13 products: base *reflectivity* (R) of particles on six elevations, *velocity* (V) on six elevation and the estimated quantity of water (VIL) contained by a one square meter column of air.

For detecting the underlying structure of the data sets *D*1 and *D*2, the SOM model (Section 2.2) is applied for obtaining an unsupervised two-dimensional representation of the data sets. The corresponding *U*-matrices will be analyzed and compared for assessing the relevance of the meteorological products used in detecting the time stamps when a certain meteorological event occurred.

5. Experimental evaluation

We are presenting in this section the experiments performed by applying the methodology introduced in Section 4 with the aim of answering RQ1 and RQ2.

5.1. Data set

This study uses data provided by the WSR-98D weather radar [16] located in Bobohalma, Romania and stored in the NEXRAD Level III format, as described in Section 3. The day used as case study is the 5th of June 2017, a day with moderate atmospheric instability manifested through thunderstorms accompanied by heavy rain and medium-size hail. In our study we selected an area from the central Transylvania region (parts of Mureş, Cluj, Alba and Sibiu counties) representing a grid having the geographical coordinates (46.076N, 46.725N, 23.540E and 25.064E). In the chosen geographical area, there were two distinct episodes with intense meteorological events in June 5, 2017: the first one between approximately 09:00 and 11:00 UTC, and the second one between approximately 12:00 and 17:00 UTC, with the most severe events taking place between 14:00 and 15:00 UTC. Concerning these phenomena, the National Meteorological Administration issued five severe weather warnings, code yellow.

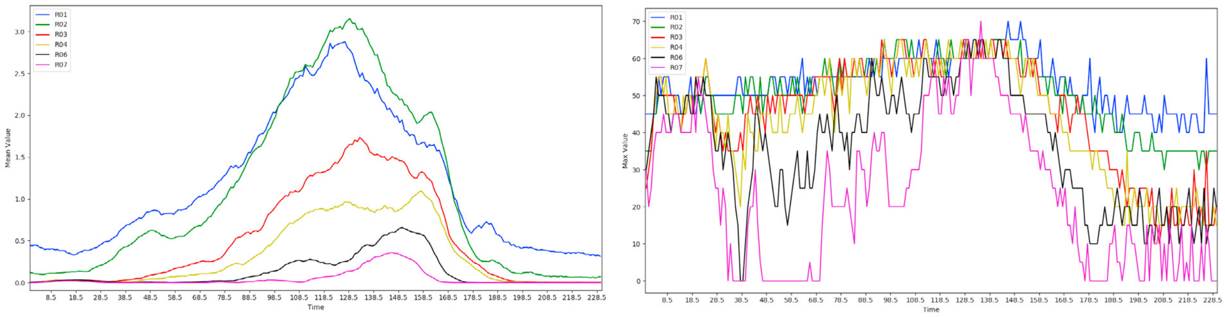
The data grid provided by the radar for the selected geographical area at a given time moment is fit to a matrix. The radar provides one data matrix for each of the 24 meteorological products, and each matrix has 624 rows and 800 columns (i.e. $m = 800$ and $n = 624$). As stated in Section 4, the radar data is split into multiple time stamps, each time stamp representing data gathered by the radar every 6 minutes (the radar takes 6 minutes to gather the data for the area). The radar data used in our case study has been recorded between 00:04:04 UTC and 23:54:02 UTC. There are missing time stamps due the fact that, as a protective measure, the radar is shut down when lightning occurs around its location. Accordingly, we have a total of 231 time stamps (i.e. $k = 231$), with time stamp 1 corresponding to 00:04:04 UTC. The most interesting time stamps are the ones in which there is data about the above mentioned meteorological events: the time stamps from 88 to 106 contain the data for the meteorological event from 09:00 to 11:00 and the time stamps from 117 to 165 contain the data for the meteorological event from 12:00 to 17:00. The data for the maximum values approximated to be between 14:00 to 15:00 are contained in the time stamps from 137 to 145.

The data gathered by the radar and exported as shown in Section 3 contains a special value that represents “No Data”. This value is usually represented by −999 but we decided to replace it with 0 as in most cases this value refers to air particles with 0 reflectivity (i.e. no significant water droplets). “No data” may also represent air volumes which have returned no signal, for example if a sector with high reflectivity is between the radar and the respective location. In this case, replacing it with 0 is also correct, since the entire region is obturated and the data is not relevant for the learning process. The radar data is prone to different type of errors, meteorological and technical, which implicitly are to be found in the output data matrix. Meteorological errors (e.g., the underestimation of a particle’s reflectivity) are difficult to identify and eliminate, but some errors occurring during the data conversion have been identified and corrected. For example, the product V should only contain values from -33 to 33 but we found values of -100. From a meteorological point of view, those erroneous values correspond to radar uncertainties in evaluating the direction and/or the speed of the particle, and are not taken into account in operative service since they are punctiform values and are irrelevant to the characteristics of a region. In order to avoid introducing into our experiments the noise that these -100 values for V represent, we decided to skip them in the unsupervised learning process. More exactly, during the training using the Kohonen algorithm, the erroneous values of -100 were omitted while computing the Euclidian distance between the input instances and the neurons from the map.

5.2. Statistical data analysis

As a preliminary step before applying the unsupervised SOM models, a statistical analysis was performed on the data set described in Section 5.1, with the goal of analyzing the variation of the meteorological products on each time stamp. For analysis, the most relevant data products from a meteorological viewpoint are used: R, V and VIL. We mention that in the data set used in our case study, values for R and V products are available for only six elevations (i.e. R01-R04 and R06-R07, V01-V04 and V06-V07). The other three elevations delivered by the radar are missing since they are not regularly used in operative service, thus they are not stored in the same format as the rest of the elevations.

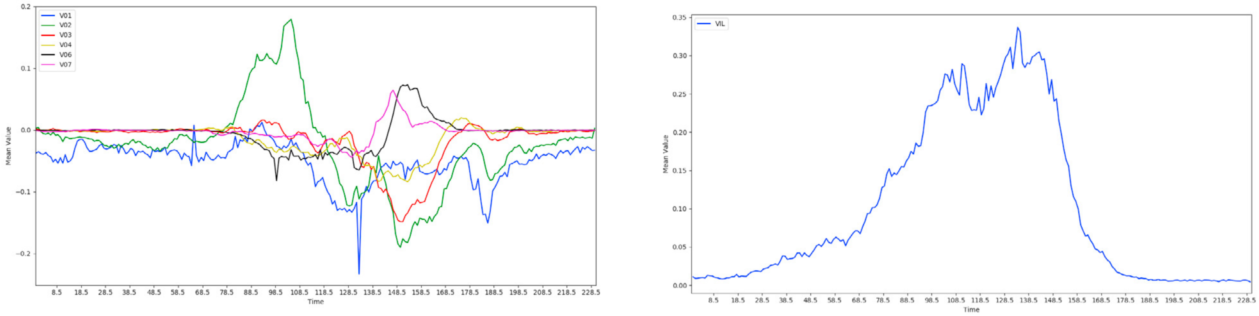
For each data product *p* (R01-R04, R06, R07, V01-V04, V06, V07 and VIL) and each time stamp *t*, we computed the average value of product *p* for all the cells from the analyzed grid. Figure 3 depicts on the leftside image the variation of the average value for a certain R with respect to each time stamp (ranging from 1 to 231). The rightside image from Figure 3 illustrates the variation of the maximum R values on each time stamp. Figure 4 depicts the histogram considering the VIL data product. On each figure, a step of 10 was selected on OX axis (i.e. an hour).



(a) Histogram for average R values.

(b) Histogram for maximum R values.

Fig. 3: Variation of average R (left) and maximum R (right) values for each time stamp.



(a) Histogram for average V values.

(b) Histogram for average VIL values.

Fig. 4: Variation of V (left) and VIL (right) average values for each time stamp.

Analyzing the histograms from Figures 3(a) and 4 we observe that from time stamp 88 (the start of the severe meteorological phenomenon), the average values for products R and VIL are increasing. Around time stamp 165 (the end of the meteorological event) the average values for all R products start to decrease. In addition, VIL average values are also significantly decreasing from time stamp 165.

Figure 3(b) does not reveal a clear correlation between the maximum R values and the severe meteorological event. The beginning of the meteorological event is undistinguishable, since the maximum R values are fluctuating before and after the start of the phenomenon. Still, we observe that the highest maximum value for R07 is reached around time stamp 138, which is in the period of maximum intensity of the severe event. From time stamp 138, the maximum R07 values show a general decreasing tendency, but there is no clear evidence of it. We also note that R02, V02 and VIL are the products which have the fastest grow from time stamp 88, with a maximum average R01 and VIL around time stamp 137 which corresponds to the maximum intensity of the meteorological event. In addition, from Figure 3(b) it can be observed that the maximum value of each R product is the highest in the period between around 137 and 160 (the highest intensity of the phenomena).

5.3. Experimental setup

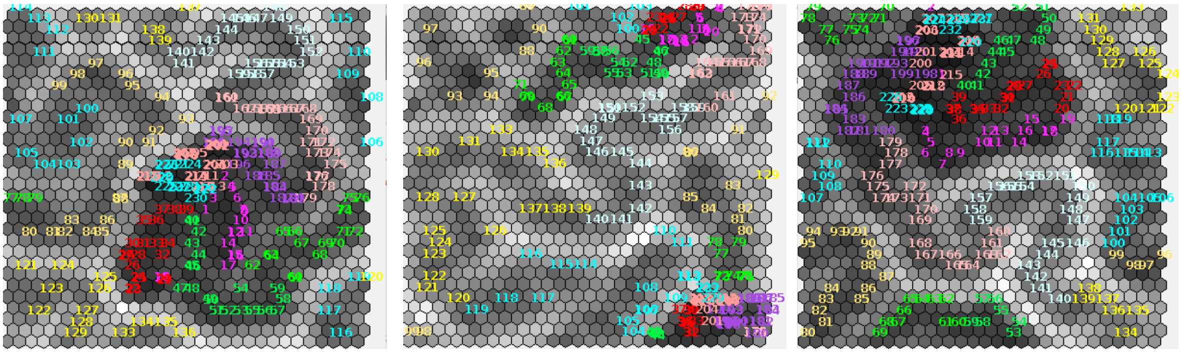
For the SOM employed in the experiments we used our own implementation, without any third party libraries. For building the SOM, we used a torus topology [8], which is shown in the literature to provide better neighborhood than the conventional lattice topology. As the distance function for discriminating between the input instances, the *Euclidean Distance* between the high-dimensional representation of instances has been adapted to ignore the erroneous values provided by the radar (in our case the values -100 for V).

The following parameters were used for the SOM: a configuration of 30x30 neurons on the map, 20000 training epochs and a learning rate of 0.1. In our implementation, the lower values on the U-matrix are depicted as darker places, while higher places are marked as whiter regions. Accordingly, darker regions encode similar data instances, whilst whiter regions represent separation boundaries between the data clusters.

6. Results and discussion

The experiment is applied following the methodology introduced in Section 4. Through this experiment, we would expect the SOM to unsupervisedly detect temporal intervals where certain meteorological phenomena occurred.

Figure 5 depicts the U-matrix visualization for the SOMs built as described in Section 5 using all 24 meteorological products (leftside image) and only 13 meteorological products (rightside image). For readability reasons, the BMUs for the instances are labeled with their corresponding time stamp and are coloured using a set of 12 colours. The multiple different colours were used in order to be easier to keep track of the time flow. Each colour corresponds to a set of 20 consecutive time stamps, thus each colour representing a 2 hour period of the total of 24 hours.



(a) SOM for all 24 meteorological products.

(b) SOM for only 13 meteorological products (R, V and VIL)

(c) SOM using only R and VIL products.

Fig. 5: U-matrix visualization for the SOMs built using the proposed data model.

Analyzing the SOMs in Figure 5 we observe that they have similar shapes and many similar elements. There is a bigger, dark, very well-defined shape (on the rightside image it is cut in half by the top/bottom margin of the image representation of the torus), which coincides very well with the time stamps for which there are no meteorological events (1 to 88 and 166 to 231). We can observe that a rope-like shape from approximately time stamp 70 to 170, in which time stamps are ordered consecutively, as in a smooth, progressive transition – this shape is more clearly delimited on the rightside image. The end of the event, time stamp 165 (and onward), is clearly delimited from other previous time stamps in the midst of the meteorological events, such as 90-95.

On both SOMs we can observe that the slow progression begins a little before the actual start of the meteorological event, at around the time stamp 70. However, on the rightside image, we can clearly distinguish an interesting phenomenon, at time stamps 71-72, which does not appear on the other image. This suggests that there might be some readable changes in the meteorological products almost 2 hours before the start of the event which might help to forecast the start of the phenomena. We also observe that on the rightside image the start of the meteorological event is clearly delimited at time stamp 88.

From what we have mentioned it appears that the SOM built using only R, V and VIL meteorological products (Figure 5b) is more accurate than the SOM from Figure 5a, as we can observe more clearly the rope-like shape. In addition, the SOM from Figure 5b reveals interesting details, such as the phenomenon at time stamps 71-72 and the clear start of the meteorological event. Consequently, we can conclude that R, V and VIL values may be useful for predicting meteorological events and that additional meteorological products (other than R, V and VIL) do not bring significant additional information about the phenomenon. Thus, instead of using all meteorological products only R, V and VIL may be enough for the nowcast of severe weather events. Still, further investigations and experiments are needed to support this conclusion.

For assessing the relevance of V in usupervisedly uncovering meteorological events, we performed the first experiment using only R and VIL products, without considering V. Thus, only 7 products were used instead of 13. The U-matrix visualization for the SOM trained using only R and VIL is illustrated in Figure 5c. Comparing the map from Figure 5c with the one from Figure 5b we observe that the start of the meteorological event, at time stamp 88, is clearly visible only in Figure 5b. In addition, the interesting phenomenon at 71-72 is not visible in Figure 5c compared to Figure 5b, thus Figure 5c is missing two distinguishing features on which we conclude that the SOM from Figure 5b is more accurate than the one from Figure 5a. Accordingly, we conclude that V is also relevant in detecting severe meteorological events and R and VIL measurements have to be used together with V for increasing the performance of the detection process.

The previous analysis of the results obtained using the first experiment confirms that the unsupervised SOM model is able to uncover severe weather phenomena in radar data.

6.1. Analysis of the results from a meteorological perspective

With the goal of answering research question RQ3, we are briefly analyzing in the following the meteorological relevance of the computational results presented in Section 6.

The weather dynamics sampled by the radar products values in the data set described in Section 5.1 is consistent with relatively small consecutive changes in the products' values occurring on a short time scale (6 minutes). In calm weather conditions these changes are typically smooth, whereas during severe meteorological events distinguishable transitions are detected. As shown in Section 5.2, the values of the meteorological products suffer significant modifications in extreme meteorological conditions and these changes are observable during the occurrence of the severe phenomenon. Concerning the radar products, *Reflectivity* (R), *particle velocity* (V) and *vertically integrated liquid* (VIL) are the main products used by meteorologists for short-term weather forecasting. The unsupervised SOM model proposed in this paper is able to capture both features, as demonstrated in Section 6: changes in the products' values from a certain geographical area occurring on short time periods are normally encoded in chained events. In addition, there is evidence that the values of the radar products clearly discriminate between calm weather and severe events. The SOM is also able to unsupervisedly detect these patterns using only R, V and VIL products. This suggests the feasibility of learning to predict (using R, V and VIL products) an entire data parallelepiped (as detailed in Section 4.1) at a certain time based on data parallelepipeds at previous time moments.

As a conclusion of our study, SOMs are able to unsupervisedly uncover in radar data patterns which are relevant from a meteorological perspective. The findings of our study suggest promising results in applying predictive supervised learning models for weather nowcasting using radar data.

7. Conclusions and further work

This paper presented a study towards applying SOMs as an unsupervised classification method for analyzing meteorological radar data and investigating the relevance of several meteorological products in detecting severe weather phenomena. Several experiments were performed on real radar data provided by the Romanian National Meteorological Administration and collected on the Central Transylvania region. Analyzing the results, both from a computational and meteorological perspective, we obtained an empirical evidence that in normal weather condition the values of the meteorological products are smoothly changing in time, excepting situations when certain severe phenomena occur. Thus, meteorological events reflected in changes occurred in the values of several meteorological products are indeed detected by unsupervised learning algorithms.

SOMs have been used in our study as self-organizing connectionist models, but further extensions of our approach will also consider alternative dynamic connectionist structures [22] which allow the dynamic analysis of data. Besides, we plan to investigate alternative unsupervised learning models such as fuzzy SOM [3], clustering [1], t-SNE [14] and *relational association rule mining* [15, 2]). Based on the study performed in this paper we aim to advance our research towards weather nowcasting based on radar data. In addition, we will investigate the appropriateness of enlarging the data set used for learning, by adding new radar data and other geographical features (i.e elevation, exposure, vegetation, hydrological and anthropic features).

Acknowledgments

The authors acknowledge the assistance received from the National Meteorological Administration from Romania, for providing the meteorological data sets used in the experiments.

References

[1] Celebi, M.E., 2014. Partitional Clustering Algorithms. Springer Publishing Company.

[2] Crivei, L.M., 2018. Incremental relational association rule mining of educational data sets. Studia Universitatis Babes-Bolyai Series Informatica 63, 102–117.

[3] Czibula, I., Czibula, G., Marian, Z., Ionescu, V.S., 2016. A novel approach using fuzzy self-organizing maps for detecting software faults. Studies in Informatics and Control 25, 207–216.

[4] Elfelly, N., Dieulot, J.Y., Borne, P., 2008. A neural approach of multimodel representation of complex processes. International Journal of Computers, Communications & Control III, 149–160.

[5] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

[6] Kalinic, H., Matic, F., Mihanovic, H., Vilibi, I., agar, N., Jasenko, B., Tudor, M., 2015. Comparison of two meteorological models using self-organizing maps, in: OCEANS 2015 - Genova, pp. 1–6. doi:10.1109/OCEANS-Genova.2015.7271374.

[7] Kaski, S., Kohonen, T., 1996. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world, in: Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets, World Scientific. pp. 498–507.

[8] Kihato, P.K., Tokutaka, H., Ohkita, M., Fujimura, K., Kotani, K., Kurozawa, Y., Maniwa, Y., 2007. Spherical and torus som approaches to metabolic syndrome evaluation., in: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (Eds.), ICONIP (2), Springer. pp. 274–284.

[9] Khler, A., Ohrnberger, M., Scherbaum, F., 2009. Unsupervised feature selection and general pattern discovery using self-organizing maps for gaining insights into the nature of seismic wavefields. Computers & Geosciences 35, 1757 – 1767.

[10] Lampinen, J., Oja, E., 1992. Clustering properties of hierarchical self-organizing maps. Journal of Mathematical Imaging and Vision 2, 261–272.

[11] Lepioufle, J.M., 2015. Towards a better use of the precipitation radar data; a feasibility study. Technical Report. Norvegian Meteorological Institute - MET.

[12] Lin, G.F., Wang, T.C., Chen;, L.H., 2016. A forecasting approach combining self-organizing map with support vector regression for reservoir inflow during typhoon periods. Advances in Meteorology , Article ID 7575126.

[13] Liu, Y., Weisberg, R.H., 2011. A review of self-organizing map applications in meteorology and oceanography, in: Mwasiagi, J.I. (Ed.), Self Organizing Maps. IntechOpen, Rijeka. chapter 13.

[14] van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605.

[15] Miholca, D.L., Czibula, G., Czibula, I.G., 2018. A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks. Information Sciences 441, 152 – 170.

[16] NOAA's National Weather Service. Radar Operations Center, 2018. NEXRAD Technical Information.

[17] Ohba, M., Kadokura, S., Nohara, D., Toyoda, Y., 2016. Rainfall downscaling of weekly ensemble forecasts using self-organising maps. Tellus A: Dynamic Meteorology and Oceanography 68, 29293.

[18] Somervuo, P., Kohonen, T., 1999. Self-organizing maps and learning vector quantization for feature sequences. Neural Processing Letters 10, 151–159.

[19] Swedish Meteorological and Hydrological Institute, 2018. Cooperation is a must for adaptation to and mitigation of climate change.

[20] Tambouratzis, T., Tambouratzis, G., 2003. Meteorological data mining employing self-organising maps, in: Pearson, D.W., Steele, N.C., Albrecht, R.F. (Eds.), Artificial Neural Nets and Genetic Algorithms, Springer Vienna, Vienna. pp. 149–153.

[21] Tambouratzis, T., Tambouratzis, G., 2008. Meteorological data analysis using self-organizing maps. Int. J. Intell. Syst. 23, 735–759.

[22] Tian, D., Liu, Y., Wei, D., 2006. A dynamic growing neural network for supervised or unsupervised learning, in: 2006 6th World Congress on Intelligent Control and Automation, pp. 2886–2890. doi:10.1109/WCICA.2006.1712893.

[23] Vilibi, I., epi, J., Mihanovi, H., Kalini, H., Cosoli, S., Janekovic, I., agar, N., Jesenko, B., Tudor, M., Dadi, V., Ivankovi, D., 2016. Self-organizing maps-based ocean currents forecasting system. Scientific Reports 6, .22924.

[24] WMO - World Meteorological Organisation, 2018. Weather Climate Water. https://www.wmo.int.