Pattern Recognition Letters 000 (2017) 1–10

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

semBnet: A semantic Bayesian network for multivariate prediction of meteorological time series data

Monidipa Das*, Soumya K. Ghosh

Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Bayesian network
Semantic similarity
Domain knowledge
Spatial semantics
Time series prediction
Meteorology

ABSTRACT

Meteorological time series prediction plays a significant role in short-term and long-term decision making in various disciplines. However, it is a challenging task involving several issues. Sometimes, the available domain knowledge may help in dealing with certain issues in this regard. This work proposes a multivariate prediction approach based on a variant of *semantic Bayesian network*, termed as *semBnet*. The key objective of *semBnet* is to incorporate the spatial semantics as a form of domain knowledge, in standard/classical Bayesian network (SBN), and thereby improving the accuracy of meteorological prediction. It has been shown that compared to SBN, the proposed *semBnet* is less prone to parameter value uncertainty. Empirical studies on multivariate prediction of *Temperature, Humidity, Rainfall* and *Soil moisture* demonstrate the superiority of proposed approach over *linear* statistical models (e.g. ARIMA, *spatio-temporal ordinary kriging* (ST-OK)), and *non-linear* prediction techniques based on ANN, SBN, *hierarchical Bayesian autoregressive* model (HBAR) etc. Most significantly, compared to SBN, the proposed *semBnet* shows average 24% improvement in *mean absolute percentage error* of prediction.

1. Introduction

The prediction of meteorological time series such as *temperature, rainfall, soil moisture, wind speed, relative humidity, atmospheric pressure* etc. plays significant role in various disciplines, including weather control, climate impact assessments, agriculture, water system management, and so on. However, the two major challenges in this regard are: 1) complex spatio-temporal interrelationships among the meteorological variables; and 2) influence of various spatial attributes, like latitude, land cover category, land elevation etc. For example, as investigated by Ding et al. [15] and Bagley et al. [2], precipitation is highly influenced by the *surface elevation* and *land-use land-cover (LULC)* type of a region. In this situation, the spatial semantics of the influencing factors can aid in prediction process by providing some added insights. For instance, the *land surface temperature* (LST) of any two places, one belonging to an *urban area* and the other belonging to a *mining region* are influenced more or less in a similar fashion (assuming all other factors to be constant) as both the locations fall under same LULC category which is *'built-up'*. On the contrary, the LST of two other locations, one at an *urban area* and the other inside an *evergreen forest* are influenced in a considerably different manner [37], since in this case the two locations belong to two different LULC categories, namely: *'built-up'* and *'forest'* respectively. Therefore, the domain knowledge on spatial semantics can play an important role in determining meteorological conditions of any location.

Down through the years, numerous models have been proposed for predicting meteorological time series. Most of these are based on various linear statistical processes such as auto-regressive moving-average (ARMA), AR integrated moving-average (ARIMA) [9,18], spatio-temporal kriging [17] etc., or based on computational intelligence (CI) techniques, like artificial neural network (ANN) [32,36], Bayesian network (BN) [1,12], support vector machine (SVM) [7], chaos theory [13] and so on.

Among these CI techniques, the Bayesian network (BN), that can intuitively represent the relevant dependencies among numerous variables, is very much suitable for multivariate prediction in meteorology [6]. With its directed acyclic graph, BN can automatically capture probabilistic information from data and can reason with uncertain knowledge [14,35]. However, one of the major problems with BN is that a proper learning of the network needs large amount of observed data be available during training. Otherwise, it may lead to strongly biased inference results with full of uncertainty [8]. It has been observed by Luo et al. [24] and Chang et al. [8] that in such case, a prior knowledge, more specifically a prior qualitative semantic knowledge, about the respective domain, may help in many ways to adjust the uncertainty. In this regard, two key objectives in our work are:

* Corresponding author.
E-mail address: monidipadas@hotmail.com (M. Das).

http://dx.doi.org/10.1016/j.patrec.2017.01.002
0167-8655/

1. Incorporation of spatial semantics in the Bayesian network model for improving the Bayesian analyses;
2. Employing this semantically enhanced Bayesian network for better modeling of spatio-temporal inter relationships among meteorological parameters and spatial attributes.

1.1. Existing works on semantically enhanced Bayesian network

Although the Bayesian networks with incorporated semantics have proved their usefulness in a number of applications, it is still not a much explored area. A few notable variants of semantic Bayesian network can be found in the works by Kim et al. [19], Butz et al. [5], Zhou et al. [38], and Madsen and Butz [25] respectively.

Kim et al. [19] have used their proposed semantic Bayesian network (SeBN) in a conversational agent to infer the detailed intentions of the user. In SeBN, the network itself contains both probabilistic and semantic relationships. The inference generation is followed by thresholding process to select the appropriate target value corresponding to the user query. Zhou et al. [38] have used semantic Bayesian network (sBN) for web mashup network construction, where sBN has been used to process all information sources on the semantic web. In order to process a semantic graph structure-based attribute, the authors have defined semantic subgraph template using a SPARQL query. The works by Butz et al. [5], and Madsen and Butz [25], are mainly on exploiting semantics in Bayesian network inference. For that purpose, Butz et al. [5] have proposed a join tree propagation architecture in which inference is conducted in a join tree (JT). Each node in JT posses a local BN that preserves all conditional independencies of the original BN. In order to use semantics in Bayesian network inference, Madsen and Butz [25] have used Lazy Propagation. It basically combines a Shenoy–Shafer propagation [26] and variable elimination scheme for computation of messages and marginals.

On the other hand, our proposed one is a new variant of semantic Bayesian network (*semBnet*) which is novel from both learning and inference generation perspectives. In order to incorporate semantics in the Bayesian analysis, the proposed *semBnet* uses a semantic hierarchy representation of the domain knowledge and some appropriate semantic similarity measures between the various concepts (refer Section 2.2). In our work, the proposed *semBnet* has been applied for better modeling of spatio-temporal inter-relationships among meteorological parameters. To the best of our knowledge this is the first attempt of using semantic Bayesian network for multivariate prediction in meteorology. However, the proposed *semBnet* is a generic model which can be applied to diverse set of applications.

1.2. Problem statement and motivations

The overall problem of meteorological time series prediction, addressed in the present work, can be stated as follows:

- Given, the historical daily time series data set over n meteorological parameters in $M = \{m_1, m_2, \ldots, m_n\}$, corresponding to a set of l locations $L = \{l_1, l_2, \ldots, l_l\}$ for previous t years: $\{y_1, y_2, \cdots, y_t\}$. Also given, the spatial attributes $SA = \left\{sa_1^l, sa_2^l, \ldots, sa_p^l\right\}$ for each location $l \in L$. The problem is to determine the daily times series of the variables in M for any location $x \in (L \cup Z)$ for future q years $\left\{y_{(t+1)}, y_{(t+2)}, \ldots, y_{(t+q)}\right\}$, when the spatial attributes of x is observed as $\left\{sa_1^x, sa_2^x, \ldots, sa_p^x\right\}$. Here, Z is a set of k new locations $\{z_1, z_2, \ldots, z_k\}$, such that $z_i \notin L$, for $i = 1$ to k, and q is a positive integer, i.e. $q \in \{1, 2, 3, \cdots\}$.

As per the definition stated above, this problem is a kind of spatio-temporal prediction that needs to consider spatial as well as temporal aspects of change in inter-relationships among the meteorological variables. Therefore, a Bayesian modeling of the problem may appear as an appropriate solution. However, challenge arises when a spatial attribute $sa \in SA$ has qualitative values with different semantic interpretations. In that case, treating such variable in a conventional manner, without utilizing the available spatial semantics, may results in improper Bayesian learning and inference. For example, consider the example scenario illustrated in Fig. 1.

Fig. 1(a) shows a causal dependency graph among three meteorological variables (*Temperature (T), Relative Humidity (H), Rainfall (R)*) and three spatial attributes (*Latitude (Lat), Elevation (Elev), LULC*), which significantly influence these meteorological variables. This graph is basically the directed acyclic graph (DAG) that forms the structure of the Bayesian network. Possible values for each of the quantitative variables (i.e. *T, H, R, Lat*, and *Elev*) are provided in terms of some discrete ranges (refer Fig. 1(b)). On the other side, *LULC* (land-use land-cover) is qualitative variable, which may take the values from its domain: {.'Urban', 'Mining', 'Forest', 'Wetland'.}. Now, suppose, for the variable LULC, some domain knowledge is also available that basically provides insights on the semantic relationships (in this case inheritance) among these domain values of LULC. This knowledge has been represented in terms of a semantic hierarchy [30] in the Fig. 2. Here, it must be made clear that this hierarchy is only the representation of the knowledge; it is not a part of the network/ causal dependency graph in Fig. 1(a). A toy data set over eight separate locations are also provided (refer Fig. 1(c)) for the variable *Temperature* (T).

In this scenario, the standard Bayesian network analyses are performed without using the domain knowledge i.e. without using the semantic relationships expressed through the hierarchy (refer Fig. 2). Therefore, as per the principles of standard/ classical BN, the probability of $T = T_3$, given $Lat = Y_1$, $Elev = E_1$, and $LULC =$ '*Urban*' becomes $P(T_3|Y_1, E_1, \text{'}Urban\text{'}) = \frac{1}{2} = 0.5$, which considers the record $< Y_1, E_1,$ '*Urban*', $T_3 >$ out of {$< Y_1, E_1,$ '*Urban*', $T_2 >$, $<Y_1, E_1,$ '*Urban*', $T_3 >$}. Thus, the standard Bayesian network treats all the domain values of LULC, like 'Urban', 'Mining' etc. as separate categorical values. However, as per the hierarchy, 'Urban' is a sub-category of LULC type 'Builtup'. Therefore, 'Urban' is somehow semantically related to 'Mining' as well. (Since the toy data set does not contain any entry on LULC type 'Rural', we have not considered it.) That means, the temperature of 'Urban' and 'Mining' area are influenced more or less in similar manner than the temperature of 'Wetland', 'Forest' etc. So, while measuring $P(T_3|Y_1, E_1,$ '*Urban*'), the effect of two more records: $< Y_1, E_1,$ '*Mining*', $T_3 >$ (corresponding to $location_1$), and $< Y_1, E_1,$ '*Mining*', $T_3 >$ (corresponding to $location_5$) in the data set should also be considered. In order to overcome such limitation in a standard/ classical Bayesian network, we've proposed a variation of semantic Bayesian network, termed as *semBnet*. The *semBnet* provides a mechanism to utilize the domain knowledge, expressed in terms of semantic hierarchical relationships, and incorporate such semantics in a standard Bayesian Analysis.

1.3. Contributions

The key contributions in the present paper can be summarized as follows:

- Defining a new variant of semantically influenced Bayesian network, termed as *semBnet*, that incorporates semantic information during probabilistic learning and inference generation
- Theoretical performance analyses of *semBnet* in comparison with the standard/ classical Bayesian network (SBN)

Please cite this article as: M. Das, S.K. Ghosh, semBnet: A semantic Bayesian network for multivariate prediction of meteorological time series data, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.01.002