International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 143-147

Spatiotemporal Data Modeling for Grid Based Subspace Clustering

G.N.V.G. Sirisha
Assistant Professor, Department of Computer Science and Engineering,
S.R.K.R. Engineering College, Bhimavaram, Andhra Pradesh, India.
E-mail: sirishagadiraju@gmail.com

M. Shashi
Professor, Department of Computer Science and Systems Engineering
A.U. College of Engineering, (Andhra University), Visakhapatnam, Andhra Pradesh, India.
E-mail: smogalla2000@yahoo.com

Abstract
This paper explains different data preprocessing techniques that are to be applied on spatiotemporal data to improve the quality and make it appropriate for grid based subspace clustering. The choice of data preprocessing techniques should be based on the nature of the data and the purpose for which it is used. Two real world datasets namely animal migration data and weather data are taken as case studies. Animal migration dataset is animal movement data annotated with weather parameters obtained by integrating the data collected from MoveBank and IRI/LDEO Climate Data Library of Columbia University. Weather dataset is the Cambridge daily weather data collected on the rooftop of University of Cambridge. Data Cleaning, data Integration, data transformation and data reduction techniques were applied on animal migration data. Data cleaning, data transformation and data reduction techniques are applied on Cambridge Weather data.

Keywords: Spatiotemporal data, multivariate time series, data preprocessing.

Introduction

Data mining is extraction of knowledge or patterns from huge amounts of data. The patterns extracted by data mining will be correct and of high quality only if the data is of high quality. Data collected in any real world database contain errors, missing values. The data required for a particular analysis task may be distributed in different databases. Due to the sophisticated data collection equipment data is easily collected and stored instantaneously as it is generated. This leads to huge amounts of data.

For effective and efficient data mining every real world dataset need to be preprocessed before applying data mining. The different data preprocessing technique are data cleaning, data integration, data transformation and data reduction [1]. Data cleaning techniques help in removing noise or errors and fill missing values. If the data that is relevant for a particular data mining task is distributed in different data sources it has to be collected from different data sources and converted to required format by data integration and data transformation techniques. Many data mining algorithms fail to mine the patterns if data size is huge. Even if an algorithm can handle huge amounts of data, it will generate huge number of patterns at high granularity and it will become difficult or impossible to comprehend the generated patterns. So data need to be converted from high granularity to low granularity using data reduction techniques.

There are a number of data cleaning, data integration, data transformation and data reduction techniques proposed in the literature. This paper argues that the selection of data preprocessing techniques should be based on the nature of data and on the type of patterns we intend to extract from the data. The paper explains the different preprocessing techniques that are to be applied on the data inorder to make it suitable for grid based subspace clustering by taking spatiotemporal data and multivariate time series (temporal data) as case studies.

Clustering aims at grouping objects such that all the objects in each group share similar characteristics. In traditional clustering all the variables/dimensions describing the objects are used in calculating the similarity between objects. The clusters that exist in subset of attributes are called as subspace clusters. A subspace cluster is defined as follows

Let $D = O \times A$ be a dataset represented in the form matrix, where **O** is the set of objects and **A** is the set of attributes. A subspace cluster **C** is a submatrix $O \times A$, where the set of objects $O \subseteq \mathbf{O}$ is homogeneous in the set of attributes defined by the subspace $A \subseteq \mathbf{A}$ [2]

Subspace clusters where members of the clusters share similar characteristics in subset of attributes could not be detected in traditional clustering. Hence many subspace clustering algorithms are proposed in the literature. Based on the approach taken these algorithms can be classified as grid based, density based and window based algorithms.

Grid based subspace clustering partitions the data space into *n*-dimensional grid. Each grid cell holds the data objects that satisfy the constraints imposed by the attribute value pairs defining the grid cell. The grid cells whose density is greater than the given threshold are called as interesting subspaces. The density of a grid cell is defined as the ratio of number of objects in the cell to the total number of objects. After identifying interesting subspaces at each dimensionality, clusters are formed by grouping dense connected subspaces.

The different data preprocessing techniques that are to be applied to make a dataset suitable for grid based subspace clustering are discussed by taking two different types of datasets as case studies. The first dataset is animal migration data taken from MoveBank which is a spatiotemporal data [3][4]. The second is climatic data collected from university

of Cambridge [5] which is a multivariate time series/temporal dataset.

Application of grid based subspace clustering on migratory animals data help us in identifying the conducive living conditions of migratory animals. Application of subspace clustering on weather data of a particular place helps us in identifying the frequently occurring weather conditions at that place. Time dimension is not required for subspace clustering so 'time' attribute is removed before preprocessing. Time attribute is required if we want to find migratory patterns of animals or seasonality in weather parameters. So time attribute can be readded if needed after subspace clustering.

Related Work

T. Mista discussed data cleaning for temporal data in the context of time series similarity computation [6]. The paper discusses different methods for filling the missing values and removing the noise. If an entire segment is missing in a time series the paper suggests not filling the missing value and ignoring that segment in both the time series in similarity computation. If the amount of missing data is less, interpolation methods are used to fill the missing value. Binning and moving average smoothing are used to remove the noise. K.R. Manjula el al. has suggested different data preprocessing techniques for multi temporal remote sensing data [7]. A novel preprocessing approach for spatiotemporal dataset was proposed by Md. Z. Islam et al. They applied the preprocessing technique on irrigation data [8]. Different time series discretization techniques were studied by P. Chaudari et al [9]. Data preprocessing and a prediction model for urban traffic in Shangai city is proposed by C.Hong et al [10].
Different preprocessing techniques to preserve seasonality in time series are studied by F. Virili et al. [11]. The paper claims that an appropriate preprocessing technique based on the properties of the time series can have a valuable effect on neural network outcome. J. Bernard et al. proposed visual interactive preprocessing of time series [12]. Though there are a number of preprocessing techniques as discussed above, methods used for preprocessing depend on data characteristics and the purpose for which we want to use the data.

Migratory Animals Data

Migratory animal's data is collected from MoveBank which is an online repository of animal tracking data. MoveBank allows users to add environmental information to animal tracking data. This work has taken data collected on the movements of "Migratory Burchell's zebra in northern Botswana". Zebras in northern Botswana make their annual migration from Okavango Delta to the Makgadikgadi grasslands. Zebras make migration to optimize their nutrition [3]. Rainfall and vegetation are highly influencing factors in migration of zebras. Subspace clustering helps in identifying frequently co-occurring rainfall and vegetation values at different places where zebras move and hence attract zebras to those locations.

The dataset used for this study consisted of 7 adult migratory zebras data. The data is recorded from October 2007 to June 2009. Zebra's location is given in terms of latitude and longitude. Animal tracking data is annotated with Weekly-Precipitation, Moderate-resolution Imaging Spectroradiometer (MODIS) Land Terra GPP 1km 8d GPP, MODIS Land Aqua GPP 1km 8d GPP, MODIS Land Terra GPP 1km 8d PsnNet, MODIS Land Aqua GPP 1km 8d PsnNet variables. 'Weekly-Precipitation' variable gives the weekly cumulative precipitation in the Tropical Rainfall Measuring Mission (TRMM) grid where the location defined by latitude and longitude is present. The rest of the variables indicate the vegetation productivity. The data set that is used for the analysis has 9 attributes which are latitude, longitude, timestamp, individual zebra identifier and the annotated weather parameters. There are 53793 such records.
Except Weekly Precipitation the data of other variables is obtained from MoveBank. The original dataset contained many other variables like individual-taxon-canonical-name, study-name, tag type etc. which were removed because they are irrelevant in identifying the conducive living conditions of migratory animals.

A. Data Aggregation

By taking the latitude and longitude values of the location where an animal is present and the time at which it is present in that location Movebank can give the values of different weather parameters in that location at that particular point of time. For example if precipitation is considered, Movebank provides 3-hourly precipitations for every *(latitude,longitude,time)* triplet. But the impact of precipitation on the climate and vegetation is long lasting hence the animal migration is influenced by weekly precipitation rather than hourly precipitation. Weekly precipitations at a place i.e. *(latitude,longitude)* pair could be computed if we can know daily precipitation at that place.
Data provided by Tropical Rainfall Measuring Mission (TRMM) from NASA (TRMM-3B42 version 7 http://mirador.gsfc. nasa.gov/) was used to find the daily rainfall at a given location. The TRMM- 3B42 algorithm combines inputs from sensors aboard multiple satellites to generate 3-hourly precipitation (mm/h) at 0.25° (~26 km) spatial resolution. The original 3-hourly TRMM rainfall estimates were converted to daily rates. TRMM gives the rainfall estimates for the spatial of coverage of 50^0 N-S and 180^0 E-W at $0.25^0 \times 0.25^0$ latitude-longitude resolution. Each $0.25^0 \times 0.25^0$ area forms a TRMM grid. The 3-hourly and daily precipitations at all TRMM grids can be obtained from IRI data library of Columbia University [13].
Zebras made their annual migration in Botswana. Botswana is an African country which is located to the south of equator and to the east of prime meridian. The places in which zebras moved are located between 23.42^0 E to 25.35^0E longitudes and 19.207^0 S to 20.811^0 S latitudes.
The trajectory in which zebras moved is covered by 24 TRMM grids. For all these 24 TRMM grids daily precipitation is collected from 1st Jan 2007 to 1st Jan 2010. The daily precipitation values are aggregated to estimate the weekly precipitation. For each TRMM grid, the weekly precipitation is obtained by adding the daily precipitations of the preceding 7 days of the week that ends with the given day.

For the calculation of Weekly-precipitation timestamp variable is also considered.

For the zebra dataset, the weekly precipitation at a location where zebra has moved is taken as weekly precipitation of the TRMM grid that encompasses the location ended by the given timestamp.

For example zebra with Id Z3864 is present in the location 19.3915^0 S latitude and 23.523^0 E longitudes on 25/10/2007. This location is present in the TRMM grid identified by its centre (19.375^0 S, 23.625^0 E). All the TRMM grids are identified by their centers. By adding/subtracting 0.125 deg. to the grid center lat/long we can find the grid boundaries. So the grid boundaries for the above grid are (19.5^0S, 19.25^0S] latitudes and (23.5^0E, 23.75^0E] longitudes respectively.
For each such TRMM grid daily precipitation is collected from 1st Jan 2007 to 1st Jan 2010. These daily precipitations are then aggregated to find the weekly precipitations. For example the weekly precipitation of the TRMM grid (19.375^0 S, 23.625^0 E) for the week that ends with 25/10/2007 is sum of daily precipitations of seven days that precede 25/10/2007. The daily precipitations of 19/10/2007 to 25/10/2007 are 0, 0, 0,0,0,0, and 0.8125331 respectively. Hence the weekly precipitation on 25/10/2007 is taken as 0.8125331.

B. Dimensionality Reduction

Zebras followed a highly directed movement during migration from Okavango Delta in the northwest to Makgadikgadi in the southeast [3]. It can be captured by applying principal component analysis to latitude and longitude. PCA helps us to identify the hidden structure of data. PCA acts both as a data transformation and data reduction technique. It is used as dimensionality reduction technique on unsupervised datasets. In this dataset PCA is applied on latitude and longitude attributes to form a new attribute. This attribute is named as lat-long. Lat-long is the first principal component obtained by applying PCA to latitude and longitude. The first principal component explained 98.3% variance in the dataset containing latitude and longitude and hence the second principal component becomes insignificant. In the context of animal migration, location though described in terms of longitude and latitude is perceived as a single dimensional attribute due to the interdependency between longitude and latitude. This natural expectation is well supported by the results of PCA.
'Prcomp' function from R Language is used for finding the principal components. R language has two functions to perform Principal Component Analysis. They are prcomp() and princomp(). princomp() performs PCA using eigenvectors. prcomp() uses singular value decomposition (SVD) to perform PCA. prcomp() has slightly better numerical accuracy, so prcomp() is generally the preferred function. princomp() will also fail if the number of variables is larger than the number of observations [14]. Latitude and Longitude are scaled and centered which normalizes them using Z-score normalization before applying PCA. Weights of longitude and latitude in the computation of the first principal component are 0.7071068 and -0.7071068 respectively. The scale and center values of latitude and longitude are given in table 1.

Table 1: Center and Scale values of Latitude and Longitude

Attribute	Center (Mean of attribute)	Scale (StdDev of attribute)
Latitude	-20.11332	0.5159210
Longitude	24.46049	0.6570401

Table 2: Loadings/ Weights by which each standardized variable should be multiplied to get transformed variable values

	Principal Component 1	Principal Component 2
Latitude Weight	-0.7071068	+0.7071068
Longitude Weight	+0.7071068	+0.7071068

The weights of the variables latitude and longitude in computing the principal components are shown in table 2.
For example consider the latitude and longitude (-19.3915, 23.52362). After applying PCA it gets transformed to (-1.997533605,-0.019003155).

These values are obtained as follows.

$$\left(\frac{latitude-center}{scale}\right)\times -0.7071068 + \left(\frac{longitude-center}{scale}\right)\times 0.7071068$$

gives the first principal component value after transformation.

$$\left(\frac{-19.391531+20.11332}{0.5159210}\right)\times -0.7071068 + \left(\frac{23.523616-24.46049}{0.6570401}\right)\times 0.7071068 = -1.997533605$$

Similarly

$$\left(\frac{latitude-center}{scale}\right)\times 0.7071068 + \left(\frac{longitude-center}{scale}\right)\times 0.7071068$$

gives the second principal component value after transformation.

$$\left(\frac{-19.391531+20.11332}{0.5159210}\right)\times 0.7071068 + \left(\frac{23.523616-24.46049}{0.6570401}\right)\times 0.7071068 = -0.019003155$$

The first principal component explained 98.3% variance in latitude and longitude variables confirming our hypothesis that latitude and longitude can be combined to form a single variable using PCA which contains the complete information about the location of zebras. So only the first principal component is retained which is named as Lat_long.
The 8 variables left after transformation are Lat-Long, timestamp, individual zebras identifier, Weekly-Prec, MODIS Land Terra GPP 1km 8d GPP, MODIS Land Aqua GPP 1km 8d GPP, MODIS Land Terra GPP 1km 8d PsnNet, MODIS Land Aqua GPP 1km 8d PsnNet. Application of subspace clustering techniques identifies the conducive living

environment of migratory animals. The analysis confines to the spatial and weather parameters observed at different timestamps irrespective of zebras identification. Hence Timestamp, individual zebras identifier which are irrelevant to subspace clustering are stripped off and the data corresponding to all the 7 zebras is taken to find subspace clusters.

C. Data Discretization

Grid based subspaces clustering algorithms work on multidimensional grids so the values of all the six variables are discretized into fixed number of intervals using equiwidth binning. The range of values of the six variables are [-2.21778,1.49096], [0,110.1495],[0,0.03395], [3.99e-009,0.022086], [-0.01248, 0.027102], [-0.00887,0.014134] respectively. The ranges of values of the variables except Weekly precipitation are discretized into 10 intervals each using equiwidth binning. Each interval of each variable is given a unique literal. The literals numbers start with 0 to 9 corresponding to Lat-long values, 10 to 19 for Weekly_Prec values, 20 to 29 for MODIS Land Terra GPP 1km 8d GPP, 30 to 39 for MODIS Land Aqua GPP 1km 8d GPP, 40 to 49 for MODIS Land Terra GPP 1km 8d PsnNet and 50 to 59 for MODIS Land Aqua GPP 1km 8d PsnNet respectively. Literal 60 is used to fill the missing values.

The values of Weekly precipitation dimension follow a skewed distribution which can be seen by the percentile information of weekly precipitation in table 3.

Table 3: Percentiles of Weekly Precipitation for Zebra Dataset

Percentile	0%	25%	50%	75%	100%
Precipitation (in mm)	0.0000000	0.0000000	0.5111257	14.7096270	110.1494600

Table 4: Percentiles of 'Sunshine' variable in 2010-2014 four hourly Weather Dataset

Percentiles	0%	25%	50%	75%	100%
Sunshine in (hrs)	0.00	0.00	0.00	0.655	16.770

Table 5: Percentiles of 'Rain' variable in 2010-2014 four hourly Weather Dataset

Percentiles	0%	25%	50%	75%	100%
Rain in (mm)	0.00	0.00	0.00	0.01	52.75

Table 6: Values used to encode wind direction

Wind Direction	E	NE	N	NW	W	SW	S	SE
Value used to encode it	1	2	3	4	5	6	7	8

D. Data Tranformation using log transformation

Inorder to make precipitation patterns more visible the weekly precipitation is made less skewed by applying $\log_2$ transformation with *log*0 set to 0. Most of the values of the attribute are 0 which is represented with a separate literal. While discretizing precipitation the first interval corresponds to 0 precipitation. The rest of the *log* transformed values are then discretized into 9 intervals using equiwidth binning. If discretization is directly applied on original data of Weekly precipitation then [0, 110.14] range is divided into 10 intervals and more than 50% of data will go into the same interval as shown in table 3.

E. Transforming the Series to Uniform Granularity

Animal tracking data is recorded periodically with different levels of granularity. While most of the data is spaced at 1 hour intervals some of the data is collected at 15 minutes intervals. For converting the data to uniform intervals data aggregation techniques are used. The dataset is thus converted to 1 hour intervals before discretization. The dataset consisted of 38491 records after applying data aggregation technique.

Weather Data

It is the Cambridge daily weather data collected on the rooftop of University of Cambridge. Temperature, Humidity, Dewpoint, Pressure, WindSpeed, WindDirection, Sun in hours, Rain, and Maximum Wind Speed are measured at half hourly intervals every day. The weather data for five years i.e. from Jan 1st 2010 to 31st Dec 2014 is taken for this research from Cambridge University. The values of all the variables are immediate at timestamp except wind speed which is average since previous timestamp and wind direction which is most frequent since previous timestamp Sun and rain values are cumulative from start of the day (i.e. they are cumulative values since 12.00AM of the same day). MxWSpd gives max wind speed since previous timestamp.

Weather data which is available at half hourly intervals is converted to hourly data using data aggregation techniques. The values of the variables Temperature, Humidity, DewPoint, Pressure and Windspeed at a given hour are taken as the average of the two values collected in that hour. WindDirection at a given hour is taken as wind direction observed at the beginning of the hour. Instead of taking cumulative values from start of day as available in original dataset, the duration of sunshine in an hour and the amount of rainfall in an hour are considered for 'Sun' and 'Rain' variables. MaxWindSpeed is maximum wind speed in that hour. Hourly data is used to find daily patterns. For finding yearly patterns in weather dataset, the hourly data is aggregated to four hourly data. Due to the skewed distribution of rain and sunshine variable values they are converted using $\log_2$ transformation. The percentile information before *log* transformation for Sunshine and rain variables in 2010-2014 weather data aggregated to 4 hours are shown in tables 4 and 5 respectively.

After data aggregation, values of the variable wind direction are encoded using integers as shown in table 6.

Every variable in the dataset except Wind direction is discretized into equal number of intervals (10) using equiwidth binning. Each interval is given a unique number. 0 to 9 correspond to Temperature, 10 to 19 correspond to Humidity, 20 to 29 correspond to Dewpoint, 30 to 39 correspond to Pressure, 40 to 49 correspond to WindSpeed, 50 to 58 correspond to WindDirection, 60 to 69 correspond to Sun in hours, 70 to 79 correspond to Rain, 80 to 89 correspond to Maximum Wind Speed. 90 is used to replace missing values. Though all missing values are replaced by a constant, it is not used in mining the patterns i.e. the missing

values are ignored in pattern extraction phase. In generation of subspace clusters ignorance of missing values does not affect the correctness of generated patterns.

Conclusions

This paper discusses different preprocessing techniques for modeling spatiotemporal data for grid based subspace clustering. Grid based subspace clustering algorithms generate subspace clusters from multidimensional grid structure. This paper has taken two real datasets as case studies to show the different preprocessing techniques that need to be applied to transform them appropriately for grid based subspace clustering for mining high quality clusters. Animal migration data is obtained from two different data sources at different time and space granularities. Data integration and data transformation techniques are applied to integrate the data. Zebras followed a highly directed movement from northwest to southeast so to capture the hidden structure of movement; PCA is applied on latitude and longitude. Weekly precipitation variable is a highly skewed variable, the skewness in the variable is reduced using *log* transformation. Most of the values of the Zebra migration are collected at hourly intervals except that a few records are collected at 15 minute intervals. The data is converted uniform granularity i.e. one hourly data using data aggregation techniques.

Pattern extraction from aggregated data is more efficient and the extracted patterns are more concise and easy to understand. So, Cambridge weather data that is present at half hourly intervals is converted to hourly intervals and four hourly intervals. The sunshine and rain variables in the dataset are highly skewed variables so *log* transformation is applied on them to reduce the skewness. Wind direction is transformed to numerical values using data encoding techniques. Grid based subspace clustering algorithms work on multidimensional grid, so all the attributes in both the datasets are discretized to form multidimensional grid using equiwidth binning.

Acknowledgment

Our sincere thanks to Hattie L.A. Bartlam-Brooks for providing access to “Migratory Burchell’s zebra in northern Botswana” data in MoveBank.

References

[1] Han, J., Kamber, M., 2006, “Data Mining Concepts and Techniques”, 2nd Edition, Morgan Kaufmann Publishers, San Francisco, CA, Chap. 2.

[2] Sim, K., Gopalkrishnan, V., Zimek, A., Cong, G., 2013, “A survey on enhanced subspace clustering”, J. Data Mining Knowl. Discov. 26(2), pp. 332–397.

[3] Bartlam-Brooks HLA., Beck PSA., Bohrer G., Harris S, 2013, “In search of greener pastures—using satellite images to predict the effects of environmental change on zebra migration”, J. Geophysical Research: Biogeosciences, 188, pp. 1–11.

[4] Bartlam-Brooks HLA., Harris S., “Data from: In search of greener pastures: using satellite images to predict the effects of environmental change on zebra migration”, Movebank Data Repository. doi:10.5441/001/1.f3550b4f

[5] Cambridge Raw Daily Weather Data**,** https://www.cl.cam.ac.uk/research/dtg/weather/index-daily-text.html

[6] Mista, T., 2010, “Temporal Data Mining”, CRC press, FL, USA, Chap.2.

[7] Manjula, K.R., Jyothi S., Anand Kumar Varma S., 2013,“Data Preprocessing in Multi Temporal Remote Sensing Data for Deforestation Analysis”, J. Global Journal of Computer Science and Technology, 13(6), pp. 19-25.

[8] Khan, Md.A., Islam Md. Z., Hafeez M., 2011,“Irrigation water demand forecasting: a data pre-processing and data mining approach based on spatio-temporal data”, Proc. AusDM conference, Vol. 121, pp.183-194.

[9] Chaudari, P., Rana, D.P., Mehta, R.G.,Mistry, N. J., Raghuvanshi, M.,M., 2014, “Discretization of Temporal Data: A Survey”, International J. of Computer Science and Information Security, 12(2), pp. 66-69.

[10] Hong, Chen., 2012, “An Urban Traffic Prediction Model Based on Temporal Data Mining in Shanghai City”, Proc. ICCEAE Yang, George., eds., AISC, 181,pp.633-639

[11] Virili, F., Freisleben, B., 1999, “Preprocessing Seasonal Time Series for Improving Neural Network Predictions”, Proc. ICSC Congress on Computational Intelligence: Methods and Applications, Soft Computing in Financial Markets, pp.622-628.

[12] Bernard, J., Ruppert, T., Goroll, O., May, T., Kohlhammer, J., 2012, “Visual-interactive preprocessing of time series data”, Proc. SIGRAD, pp. 39-48

[13] NASA GES-DAAC TRMM_L3 TRMM_3B42 v7 daily Surface Rain from all Satellite and Surface data tables. http://iridl.ldeo.columbia.edu/sources /.nasa/ .ges-aac/.trmm_l3/.trmm_3b42/.v7/. Daily/precipitation/datatables.html

[14] Mankin, E., “Principal Components Analysis A How-to Manual for R”, http://people.tamu.edu/~ alawing /materials/ESSM689/pca.pdf