ISSN 2347-3487

ASSOCIATION RULE MINING ON METROLOGICAL AND REMOTE SENSING DATA WITH WEKA TOOL

ANIL RAJPUT[1],P. K. PUROHIT[2], LL DUBEY[3], RAJESH SHARMA[4] AND RAMESH PRASAD AHARWAL[5]

[1]Department of Mathematics CSA, Govt. P. G. College Sehore (M.P.) , India
[2]National Institute of Technical Teachers' Training and Research, Bhopal, India
[3]Govt. Mahatma Gandhi Memorial P G College, Itarsi
[4]Department of Physics, SHREE Institute of Sci. and Tech., Bhopal, India
[5]Department of Mathematics and Computer Sci., Govt. P.G. College, Bareli (M.P.), India
[1]E-mail: drar1234@yahoo.com
[2]E-mail: purohit_pk2004@yahoo.com
[4]Email: rsrsharma288@gmail.com
[5]Email: ramesh_ahirwal_neetu@yahoo.com

ABSTRACT

Drought is one of the major environmental disasters in many parts of the world. There are several possibilities of drought monitoring based on ground measurements, hydrological, climatologically and Remote Sensing data. Drought indices that derived by meteorological data and Remote Sensing data have coarse spatial and temporal resolution. Because of the spatial and temporal variability and multiple impacts of droughts, we need to improve the tools and data available for mapping and monitoring this phenomenon on all scales. In this paper we present discovering knowledge by association rules from metrological and Remote Sensing data and we have also used descriptive modeling. For calculating drought taking metrological data which is extract from metrological department of Pune at Maharastra (India) and Remote Sensing data is extract from National Aeronautics and Space Administration (NASA).

Council for Innovative Research

Peer Review Research Publishing System

Journal of Advances in Physics

Vol 3, No.1

editor@cirworld.com

www.cirworld.com, member.cirworld.com

ISSN 2347-3487

1. Introduction

Basically a drought index is a function consisted of different drought-based environmental factors, resulted as a number at the end. Standardized Precipitation Index SPI and Normalized Difference Vegetation Index (NDVI) are one of the most important indices upon which drought evaluation is achieved with the range of precipitation. Upon this index, drought severity is determined regarding the related classification. Some researchers use different classifications with the understudy climate conditions. Data mining is the core task in the process known as Knowledge Discovery in databases. It consists of applying computational techniques to extract useful pattern or knowledge from the given data, typically expressed in the form of a predictive or descriptive model. Knowledge discovery in databases (KDD) was initially defined as:

"KDD is the non-trivial extraction of implicit, previously unknown and potentially useful information from data".

1.1 Knowledge Discovery and Database (KDD)

"KDD is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern or knowledge in data" Fayyad et. al. ([5], 1996) stated that data mining is a step in the KDD process concerned with applying computational techniques to actually finding patterns in the data.

2. Methods and Elements of Drought:

In this section we are discussed various elements, methods and factor which are used to calculate or monitoring drought conditions.

2.1 Earth Skin temperature (T_{skin})

In the year 1997 Jin and others used the term `skin temperature'. It has been used for `radiometric surface temperature' Jin et al ([8], 1997). It can be measured by either a hand-held or aircraft-mounted radiation thermometer, as derived from upward long wave radiation based on the Stefan–Boltzmann law, Holmes([7], 1969), Oke ([12], 1987), or retrieved from satellite observations and mapped over large areas, after removing the effects of atmospheric attenuation on satellite-measured radiances, Saunders([15], 1967), Anding and Kauth ([2], 1970), Sobrino et al ([16], 1994), Stephens ([17], 1994), Ulivieri et al ([18], 1994), Prata ([13], 1993). The retrieval techniques for finding Tskin from satellite measurements for land applications have developed substantially in the last two and half decades, Price([14], 1984). The accuracy of many Tskin algorithms has been improved to within 0.5–1°C for field measurements and for satellite observations under clear sky conditions, Becker and Li ([3], 1995), Wan and Dozier ([19], 1996), Coll et al ([4], 1994).

The definition of Tskin in satellite remote sensing is based on Planck's law. Radiances measured by satellite sensors observing the earth surface consist of both atmospheric and land surface radiation. The surface radiation at a specific wavelength gives the brightness surface temperature, defined as:

$$T_{skin}^{b} = B_{\lambda}^{-1}(L\uparrow) \qquad \text{(i)}$$

where $L\uparrow$ is the radiance measured by the radiometer, after corrections for atmospheric and emissivity effects, and B – 1 is the inverse of Planck's law Bλ(T). Planck's law gives dependence on temperature of the radiative emission from a black body, i.e., surface spectral emissivity ελ = 1. Since the land surface is typically not a black body, equation (i) needs to be modified to account for variations in ελ, e.g., Li and Becker ([11], 1993):

$$T_{skin}^{b} = B^{-1}[(L\uparrow - (1-\varepsilon_{\lambda})L\downarrow)/\varepsilon_{\lambda}] \qquad \text{(ii)}$$

where $L\downarrow$ is the downward radiance from the atmosphere.

2.2 Cloud Optical Thickness or Cloud Optical Depth(δc):

It is a measure of attenuation of the light passing through the atmosphere due to the scattering and absorption by cloud dropelts.

The cloud optical depth or optical thickness(δc) is defined as the integrated extinction coefficient over a vertical column of unit cross section. Extinction coefficient is the fractional depletion of radiance per unit path length (also called attenuation especially in reference to radar frequencies). Following are main applications of Cloud Optical Thickness.

(1) Radiative Transfer Model

(2) Earth Radiation Budget

(3) Climate Changes

Formula for calculation of Cloud Optical Thickness is

$$\delta_c = \int_0^{\infty}\int_0^{\infty} n(r)Q_e\pi\, r^2 dr\, dz \approx 2\pi \cdot \int_0^{\infty}\int_0^{\infty} n(r)r^2 dr\, dz\,. \qquad \text{(iii)}$$

The extinction efficiency depends on droplet radius r, wavelength λ and refractive index of water or ice m and is defined as the ratio of the extinction to the cross-sectional areas of droplets. It can be derived by applying the Mie theory, assuming spherical particles. The size spectrum of the cloud particles is denoted with n(r) and z is the vertical coordinate. Qe tends to become a constant value

2.3 CO_2 Fraction

Carbon dioxide (CO_2) fraction indicates the amount of CO2 in the atmosphere by volume. It is usually expressed in units of parts per million (ppm), i.e., it is the number of molecules of CO2 represented in a million atmospheric gas (air) molecules. Currently, the average atmospheric CO2 fraction is about 380 ppm (380/1,000,000 or .00038), and it varies around the world by about ± 5 ppm.

Measurement: CO2 fraction is calculated by observing the absorption of energy by CO2 molecules for a specific wavelength of radiation. Observational frequency is about twice a day.

2.4 Enhanced Vegetation Index

The Enhanced Vegetation Index (EVI) was developed as an alternative vegetation index to address some of the limitations of the NDVI. The EVI was specifically developed to:

be more sensitive to changes in areas having high biomass (a serious shortcoming of NDVI),

reduce the influence of atmospheric conditions on vegetation index values, and

correct for canopy background signals.

EVI tends to be more sensitive to plant canopy differences like leaf area index (LAI), canopy structure, and plant phenology and stress than does NDVI which generally responds just to the amount of chlorophyll present. With the launch of the MODIS sensors, NASA adopted EVI as a standard MODIS product that is distributed by the USGS (see below).

EVI is calculated as

$$EVI = 2.5 * \frac{(NIR - RED)}{(NIR + C_1 * RED - C_2 * BLUE + L)} \quad \text{(iv)}$$

where NIR, RED, and BLUE are atmospherically-corrected (or partially atmospherically-corrected) surface reflectances, and C1, C2, and L are coefficients to correct for atmospheric condition (i.e., aerosol resistance). For the standard MODIS EVI product, L=1, C1=6, and C2=7.5.

2.5 Standardized Precipitation Index (SPI)

The SPI was developed by McKee et al. ([9], 1993; [10], 1995) and defined as the number of standard deviations that the observed cumulative rainfall at a given time scale would deviate from the long-term mean. The SPI was designed to quantify the precipitation shortfall for multiple time scales. The SPI calculation for any location is based on the long-term precipitation record for a desired period. This long-term record is fitted to a probability distribution, which is then transformed into a normal distribution so that the mean SPI for the location and desired period is zero. Positive SPI values indicate greater than median precipitation, and negative values indicate less than median precipitation.

3. Data Mining

Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.
Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data, Han and Kamber, ([6], 2006). Essentially, the basic tasks of data mining and KD are to extract particular information from existing databases and convert it into understandable or sensible conclusions or knowledge. As indicated above, data mining can be viewed as a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data.

3.1 Association Rule

Association is a data mining task that discovers relationships among data. The outcome of association is association rules, showing attribute-value conditions that occur frequently together in the data set. Association rule induction model deals with the task of finding an association rule between two set of items in the expression of X → Y, Agrawal et al. ([1], 1996). The intuitive meaning of such a rule is that transactions in a database, which contain X, tend to contain Y as well. In association rules, we use the confidence and support of the rule to induce set of association rules. The task is to find all association rules that satisfy minimum support and minimum confidence constraints. The confidence of the rule is calculated as the ratio of, NXY, the number of records having true values for all items in X and Y to, NX, the number of records having true values for all items in X. The support of the rule is the ratio of, NXY, the number of records having true values for all items in X and Y to the number of all records in the database, ND. Support:

ISSN 2347-3487

3.2 Data Source and Data Description with Study Area

In our study we have included Bhopal division. Bhopal, the capital of Madhya Pradesh, is a charming blend of scenic beauty, old historic city and modern urban planning. It is located around two artificial lakes amidst picturesque settings and is also known as the city of lakes. It is the 11th century city Bhojpal, founded by Raja Bhoj, but the present city was established by an Afghan soldier, Dost Mohammed (1707-1740). His descendants built Bhopal into a beautiful city The geographical location of the Bhopal City lies within North Latitude 23°16' and East Longitude 77°36'. The location of Bhopal falls in the northwestern portion of Madhya Pradesh. If seen in the Map of India, Bhopal occupies the central most region of the country. The city of Bhopal shares its border with two large and picturesque lakes. Like few other big cities of the country**.**

3.3 Data sets description given as in following table.

S.No.	Attribute	Value/classes
1	Year	2002-2012
2	Cloud_Optical Depth	High, Normal, low
3	surface_skin_temp	High,Normal
4	CO2	High, Normal
5	EVI	Low, Very low, high, Extremely high
6	SPI	High, Normal, low, Very low
7	Drought_status	Wet, Dry, Normal

Table 1 Dataset Description

3.4 Software Issue: WEKA

Waikato Environment for Knowledge Analysis, called shortly WEKA, is a set of state-of-the art data mining algorithms and tools to in-depth analyses. The author of this environment is University of Waikato in New Zealand. The programming language of WEKA is Java and its distribution is based on GNU General Public License. These complex algorithms may be applied to data set in the aim of detailed analyses and evaluation of data mining examination. There are three main ways of WEKA use. First is analyzing data mining methods' outputs to learn more about the data; next is generation of model for prediction of new instances and finally the last comparison of data mining methods in order to chose the best one as a predictor. WEKA consists of four user interfaces out of which three are graphical and one command line. The main interface is called Explorer. It is graphical interface built of menu section and six panels connected to various data mining methods. It enables data pre-processing, classification, cauterization, and mining associations among attributes. Furthermore there is a possibility to selected attributes with the attribute evaluator and search method. The last option is visualization plotting the dependencies among attributes. The next graphical interface, Knowledge Flow is dedicated to selecting components from the tool bar and placing them on the special canvas, connecting them into directed graph than Processing and analyzing. Furthermore the data stream data processing can be designed and executed with the usage of this interface. To compare performance of data mining algorithms it is useful to chose third graphical interface called Experimenter. This module allows one to evaluate how well various data mining Methods perform for given datasets. This process is automated and statistics can be saved. This module is a most important part of the experiment. It makes in-depth statistics which are useful in case of medical datasets. After the selection of various methods, their parameters and datasets, it is possible to prepare statistic which are priceless in case of medical diagnosis support. Experimenter and Explorer are two mainly used interfaces in this thesis. WEKA allows analyzing the data sets saved in the .arff files.

3.5 Data Preparation

The process of data cleaning and preparation is highly dependent on the specific data mining algorithm and software chosen for the data mining task. The researcher attempted to prepare the data according to the requirements of the selected data mining software, Weka and selected data

mining algorithm, apriori. Weka is multi-functional data mining software. The major data mining

functions incorporated in the software are data preprocessing, classification, association, clustering and visualizing input and output. Apriori is the only association rule algorithm implemented in Weka.

ISSN 2347-3487

4. Experimental screen shots

Figure 1 weka explorer

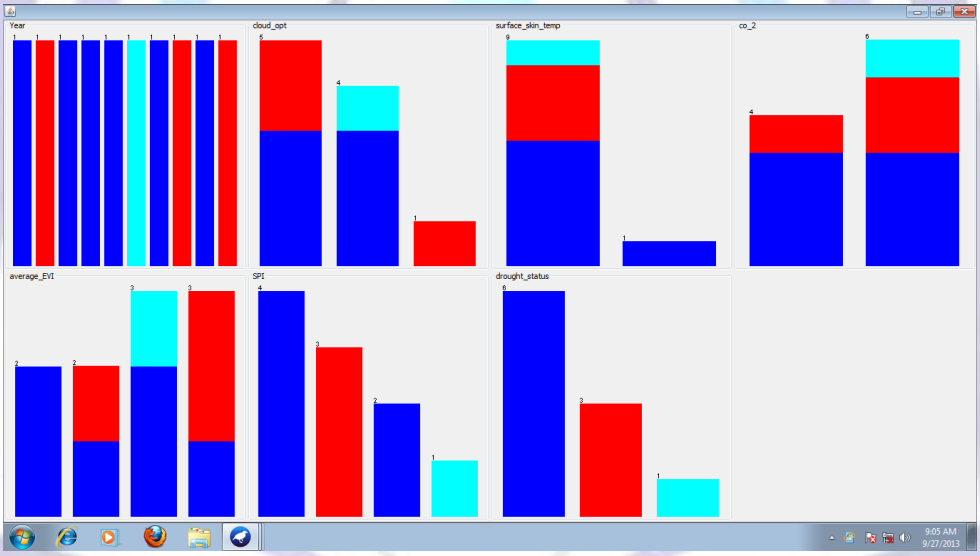

Figure 2 attribute values visualize

Viewer

Relation: drought1data1

No.	1: Year Nominal	2: cloud_opt Nominal	3: surface_skin_temp Nominal	4: co_2 Nominal	5: average_EVI Nominal	6: SPI Nominal	7: **drought_status** Nominal
1	2002-...	high	high	normal	low	low	normal
2	2003-...	high	high	normal	very low	very low	dry
3	2004-...	high	high	normal	very low	high	normal
4	2005-...	normal	normal	normal	high	low	normal
5	2006-...	high	high	high	high	high	normal
6	2007-...	normal	high	high	high	very h...	wet
7	2008-...	normal	high	high	low	low	normal
8	2009-...	high	high	high	extremly high	very low	dry
9	2010-...	normal	high	high	extremly high	low	normal
10	2011-...	low	high	high	extremly high	very low	dry

Figure 3 data set viewers

ISSN 2347-3487

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: drought1data1

Instances: 10

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.35 (3 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 19

Size of set of large itemsets L(3): 8

4.1 WEKA Generated Association Rules

S.No.	Association rule found	confidence	lift
1	co_2=high 6 ==> surface_skin_temp=high 6	100%	1.11
2	cloud_opt=high 5 ==> surface_skin_temp=high 5	100%	1.11
3	SPI=low 4 ==> drought_status=normal 4	100%	1.67
4	average_EVI=extremly high 3 ==> surface_skin_temp=high 3	100%	1.11
5	SPI=very low 3 ==> surface_skin_temp=high 3	100%	1.11
6	drought_status=dry 3 ==> surface_skin_temp=high 3	100%	1.11
7	average_EVI=extremly high 3 ==> co_2=high 3	100%	1.67
8	average_EVI=extremly high 3 ==> co_2=high 3	100%	3.33
9	SPI=very low 3 ==> drought_status=dry 3	100%	3.33
10	surface_skin_temp=high co_2=normal 3 ==> cloud_opt=high 3	100%	2
11	cloud_opt=high co_2=normal 3 ==> surface_skin_temp=high 3	100%	1.11
12	cloud_opt=high drought_status=normal 3 ==> surface_skin_temp=high 3	100%	1.11
13	cloud_opt=normal co_2=high 3 ==> surface_skin_temp=high 3	100%	1.11
14	cloud_opt=normal surface_skin_temp=high 3 ==> co_2=high 3	100%	1.67
15	cloud_opt=normal drought_status=normal 3 ==> SPI=low 3	100%	2.5
16	cloud_opt=normal SPI=low 3 ==> drought_status=normal 3	100%	1.67
17	co_2=high average_EVI=extremly high 3 ==> surface_skin_temp=high 3	100%	1.11
18	surface_skin_temp=high average_EVI=extremly high 3 ==> co_2=high 3	100%	1.67
19	average_EVI=extremly high 3 ==> surface_skin_temp=high co_2=high 3	100%	1.67
20	co_2=high drought_status=normal 3 ==> surface_skin_temp=high 3	100%	1.11

Table 2 generated association rules

5. Conclusion

Association rule is a descriptive data mining technique. This paper presents the association rule learning method from metrological drought data with Weka data mining software. We have found 20 best association rules which are shown in table 2. Each rules have 100% confidence but lift chart values are different. In table 2, highlighted rules interpreted as rule no. 9 “SPI is very low then drought status is dry is 100% same thing rule no 6, drought status is dry then surface_skin_temp is high in 100% cases.

6. References

1. Agrawal R. and Shafer J. C. (1996): Parallel Mining of Association Rules, IEEE Transactions on Knowledge and Data Engineering. 8(6), 310-316.
2. Anding D. and Kauth R. (1970): Estimation of sea surface temperature from space, Remote Sens. Environ. 1, 217–220.
3. Becker F. and Li Z. (1995): Surface temperature and emissivity at various scales: definition, measurement and related problems, Remote Sens. Rev. 12, 225–253.
4. Coll C., Caselles V. and Schmugge T. J. (1994): Estimation of land surface emissivity differences in the split-window channels of AVHRR, Remote Sens. Environ. 48, 127–134.
5. Fayyad U., Piatetsky-Shapiro G. and Smyth P. (1996): Advances in Knowledge discovery and data mining, AAAI/MIT press.
6. Han J. and Kamber M. (2006): Data Mining: Concept and Techniques, Morgan Kaufmann.
7. Holmes R. M. (1969): Airborne measurements of thermal discontinuities in the lowest layers of the atmosphere, Paper Presented at 9th Conf. Agric. Meteorol. (Seattle, WA) p 18.
8. Jin M., Dickinson R. E. and Vogelmann A. M. (1997): A comparison of CCM2/BATS skin temperature and surface-air temperature with satellite and surface observations, J. Clim. 10, 1505–1524.
9. McKee T. B., Doesken N. J. and Kleist J. (1993): The relation of drought frequency and duration to time scales, Proceedings of the Eighth Conference on Applied Climatology, American Meteorological Society, Boston. pp. 179–184.
10. McKee T. B., Doesken N. J. and Kleist J. (1995): Drought monitoring with multiple time scales, Proceedings of the Ninth Conference on Applied Climatology, American Meteorological Society, Boston. pp. 233–236.
11. Li Z. L. and Becker F. (1993): Feasibility of land surface temperature and emissivity determination from AVHRR data, Remote Sens. Environ. 43, 67–86.
12. Oke T. R. (1987): Boundary Layer Climates 2nd Edn. (London: Methuen)
13. Prata A. J. (1993): Land surface temperatures derived from the advanced very high resolution radiometer and the along-track scanning radiometer, 1. Theory J. Geophys. Res. 98 (16), 689–702.
14. Price J. C. (1984): Land surface temperature from measurements from the split window channels of the NOAA 7 advanced very high resolution radiometer J. Geophys. Res. 89, 7231–7237.
15. Saunders P. M. (1967): Aerial measurements of sea surface temperature in the infrared, J. Geophys. Res. 72, 4109–4117.
16. Sobrino J. A., Li Z. L., Stoll M. P. and Becker F. (1994): Improvements in the split-window technique for land surface temperature determination, IEEE Trans. Geosci. Remote Sens. Lett. 32, 243–253.
17. Stephens G. L. (1994): Remote Sensing of the Lower Atmosphere, An Introduction (Oxford: Oxford University Press)
18. Ulivieri C., Castronuovo M. M., Francioni R. and Cardillo A. (1994): A split window algorithm for estimating land surface temperature from satellites, Adv. Space Res. 14, 59–65.
19. Wan Z. and Dozier J. (1996): A generalized split-window algorithm for retrieving land-surface temperature from space, IEEE Trans. Geosci. Remote Sens. Lett. 34, 892–904.