International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:6, 2015

Spatio-Temporal Data Mining with Association Rules for Lake Van

T. Aydin, M. F. Alaeddinoglu

Abstract**—**People, throughout the history, have made estimates and inferences about the future by using their past experiences. Developing information technologies and the improvements in the database management systems make it possible to extract useful information from knowledge in hand for the strategic decisions. Therefore, different methods have been developed. Data mining by association rules learning is one of such methods. Apriori algorithm, one of the well-known association rules learning algorithms, is not commonly used in spatio-temporal data sets. However, it is possible to embed time and space features into the data sets and make Apriori algorithm a suitable data mining technique for learning spatio-temporal association rules. Lake Van, the largest lake of Turkey, is a closed basin. This feature causes the volume of the lake to increase or decrease as a result of change in water amount it holds. In this study, evaporation, humidity, lake altitude, amount of rainfall and temperature parameters recorded in Lake Van region throughout the years are used by the Apriori algorithm and a spatio-temporal data mining application is developed to identify overflows and newly-formed soil regions (underflows) occurring in the coastal parts of Lake Van. Identifying possible reasons of overflows and underflows may be used to alert the experts to take precautions and make the necessary investments.

Keywords**—**Apriori algorithm, association rules, data mining, spatio-temporal data.

I. Introduction

PEOPLE, throughout the history, have made inferences and future predictions based on past experiences. Nowadays, foresights about the future have gained much importance, especially when planning a business investment. The investment plans heavily benefit from the statistical and learning models constructed from the past data. Business and government agencies are continuously increasing their budgets to employ more powerful database management systems and store as much data as possible on these systems [1].

Computer technologies are widely used to forecast about the future. Software employing artificial intelligence, expert systems and data mining technologies evaluate past data (also called as training data) and develop forecasting models [2]. In this study, we developed a software employing association rules learning, a well-known data mining technique.

Data mining is the process of obtaining valuable information among huge amounts of raw data. Data mining techniques find the previously hidden relations existing in the data and describe these relations in the form of classification rules, association rules, clusters etc. These relations are later used to make future predictions on unseen data [3], [4].

As we dealt with association rules, Apriori algorithm [5], one of the widely used association rules learning algorithms, was used in our study. It is a pity that Apriori algorithm is not widely used in data sets involving time and/or space dependent features. However, it is possible to embed time and space features into the data sets and make Apriori algorithm a suitable data mining technique for learning spatio-temporal association rules.

Lake Van, the largest lake of Turkey and the subject of this study, is a closed basin. This property of the lake causes the volume of the lake to increase or decrease as a result of change in water amount it holds. In this study, lake altitude, amount of rainfall, evaporation on the surface of the lake, humidity and temperature parameters recorded in Lake Van region throughout the years are used by the Apriori algorithm and a spatio-temporal data mining application is developed to identify water-overflows and underflows (resulting in newly-formed soil regions) occurring in the coastal parts of Lake Van. Identifying water-overflows and underflows may alert the experts to take precautions and make the necessary investments.

Lake Van is located on the western part of the 16096 km^2-width basin. It has a surface area of 2626 km^2 and a drainage area of 12470 km^2, respectively. Its volume is 607 km^3 and has a maximum depth of 451 meters. These properties make it the fifth biggest lake in the world. Furthermore, it is the biggest lake in Turkey. Lake Van is a closed basin. Therefore, it sometimes overflows as a result of increase in the amount of water it holds. On the other hand, in case of heavy usage of the lake water, its volume decreases substantially. That is, its water amount has a dynamic property. In case of overflows, agricultural lands and city center may submerge. Conversely, when people heavily use lake water and/or because of changes in the climate of the region, shallow parts of the lake drain and new land formations are observed in coastal parts of the lake.

Throughout the years, changes in the water level (especially successive rapid rises) have given great damage to the agricultural lands, roads, privately held and public settlements found near the coastal regions of the lake. Even a one-meter rise in the water level can result the coast line to move tens of meters towards the land in regions where slope and altitude parameters are low [6].

We can take precaution and make necessary investments to prevent damage of lands found in the coastal parts of the lake so that scenes like in Fig. 1 will not be witnessed anymore.

T.Aydin is with the Ataturk University, Faculty of Engineering, Department of Computer Engineering, 25240, Erzurum, Turkey (phone: +90 5327161892; Fax :+904422312766; e-mail: atolga@atauni.edu.tr).

M.F.Alaeddinoglu is with the Ataturk University, Open Education Faculty, 25240, Erzurum, Turkey (e-mail: f.alaeddinoglu@atauni.edu.tr).

Also, academic studies should be employed to learn the causes of the possible changes in water level of the lake.

In our study, we learned association rules among past meteorological data and used these rules in combination with varying geographical coordinates of Lake Van itself and some specified regions around it both to forecast and validate possible overflows and underflows. The meteorological data parameters are as follows: lake altitude, amount of rainfall, evaporation on the surface of the lake, humidity and temperature. We collected data about these parameters from 22 meteorological stations recorded during 1990-2011 period. Some of these stations are: Adilcevaz, Ahlat, Çaldıran, Erçek, Erciş, Gevaş, Göldüzü, Güzelsu, Muradiye, Ovakışla, Özalp, Bitlis Reşadiye, Tatvan and Van stations. The relationships between these parameters were presented in the form of association rules. Finaly, we analysed these rules to find possible causes of over and underflows.

Fig. 1 Houses lying in the water as a result of overflow on the Iskele district of Lake Van

II. Materials And Methods

General Directorate of Meteorology keeps several meteorological data about Lake Van on a regular basis. On the other hand, in the case of retrieving geographical information, we had two choices: Using maps prepared by General Command of Mapping, or relying on online maps.

We used online maps provided by Bing search engine of Microsoft to obtain location information of borders of Lake Van and borders of possible overflow-underflow areas. Online maps are generally prepared once a year, whereas new maps by General Command of Mapping are prepared every five years. Furthermore, online maps show more details. On a scale of 1/1000, an online map shows location information up to one meter sensitivity.

Fig. 2 shows the map of Lake Van constructed by plotting 3000 border location points. Each point is represented by an ordered pair of meridian and parallel values.

Fig. 3 shows 28 possible overflow-underflow areas, each of which is constructed by plotting about 500 border location points. This figure also tells us the direction of overflows and underflows in related overflow-underflow areas.

For a specific geographical region (lake itself or any overflow-underflow area), we prepared five different maps at altitudes 1647, 1648.3, 1649.3, 1650.2 and 1651 meters. We determine the degree of overflow or underflow events byintersecting the related maps of lake and the overflow-underflow areas.

MS-SQL Server 2008 R2 database program provides us many data types to use when describing geographical regions. We employed "Multipolygon" data type to describe Lake Van and 28 overflow-underflow areas. We show how to represent Lake Van in the following example:

```
Example: INSERT INTO [regionDB].[dbo].[lakeVanTB]
([id],[name],[coordinates])
VALUES
('7','LakeVan',geography::STGeomFromText('MULTIPOLYGON
(((42.91397094726564 38.4240080236828, 43.58413696289062
38.91240739487224, 43.582763671874986 38.91187310908562,
43.58207702636717 38.91187310908562,
.
.
.
42.91397094726564 38.4240080236828 )))',4326));
```

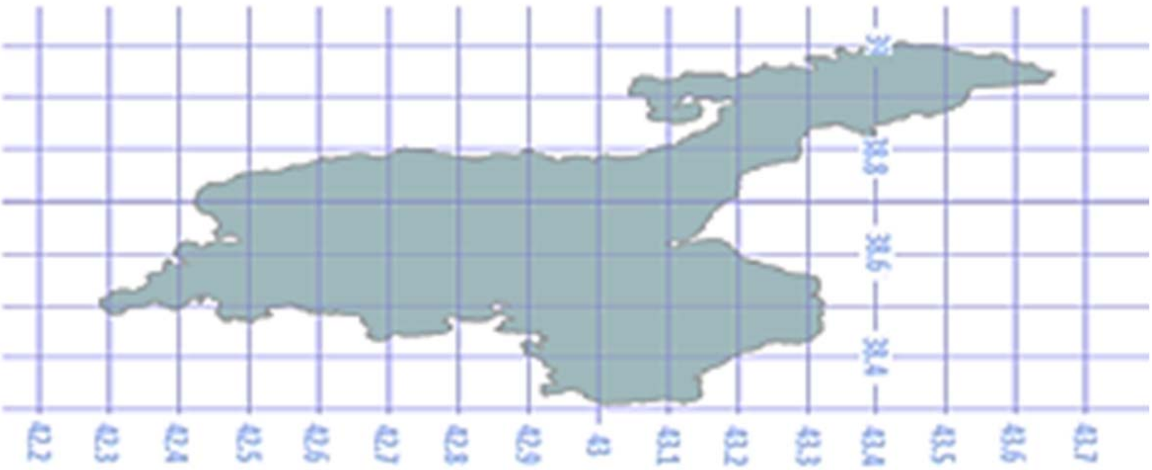

Fig. 2 Location information of Lake Van

Fig. 3 Overflow areas and direction of overflows

Fig. 4 shows the user-interface of the software developed for the study. While preparing the data set, we used the meteorological data collected from 22 stations during 1990-2011 period. That is, for each date, we collected meteorological data from all available stations. For a specific geographical region (lake itself or any overflow-underflow area), as mentioned in above paragraphs, we prepared five different maps for altitudes 1647, 1648.3, 1649.3, 1650.2 and 1651 meters. While drawing a map at a specific altitude, we employed the online maps served by Bing search engine and obtained the related meridian and parallel values of the border points of the geographical regions. We determined the degree of overflow or underflow events by intersecting the related maps of Lake Van and the overflow-underflow areas.

Part 1 of Fig. 4 serves us general menus related to the software. Some of them are: File, Apply Association Rule Mining, Lake Info and About.

Part 2 gives us lake altitude, amount of rainfall, evaporation on the surface of the lake, humidity and temperature values measured at a specific date by the given station. The specific date is chosen at Part 5, 6 and 7, whereas the station name is chosen at Part 8. The related meteorological parameter values are presented at Part 10, upon pressing the “Show” button located in Part 9.

Part 3 shows the association rules induced from the data set. While applying Apriori algorithm, minimum support and minimum confidence thresholds were selected as 20% and 25%, respectively. To be more precise, Part 3 can be analysed as follows: By pressing the “Show Rules” button located in Part 11, association rules are shown in Part 12. Each rule is presented with its corresponding confidence (accuracy) value. There are 22 stations and a set of association rules for each year between 1990-2011 is induced for each station. The number of such sets is 22 x 22 = 484. For any association rule “r” to be presented in Part 3, we require “r” to be induced at least 3 times by each of at least 3 different stations. For the sake of simplicity, if we assume “r” to be induced 3 times by each of at least 3 different stations, any station need not induce “r” for 3 successive years. Any 3 years among 1990-2011 will suffice for each of at least 3 stations. This requirement is put to ensure statistical strength of “r”. Some of the rules presented in Part 12 may be more interesting than the others. Therefore, by pressing “Select the Rule” button in Part 13, the user may select the rules he/she is interested in and those interesting rules may be shown in Part 14. This ensures that our software not only computes objectively interesting rules but also eliminates subjectively uninteresting ones among the objectively interesting rules. The user may also be interested in analyzing frequent item-sets including some items that he/she is concerned with. In this situation, he/she can select the item-set in Part 15, and press the “Show item-sets” button located in Part 16 to see the whole item-sets that include the item-set selected in Part 15. Each related item-set is presented along with its corresponding support value in Part 17.

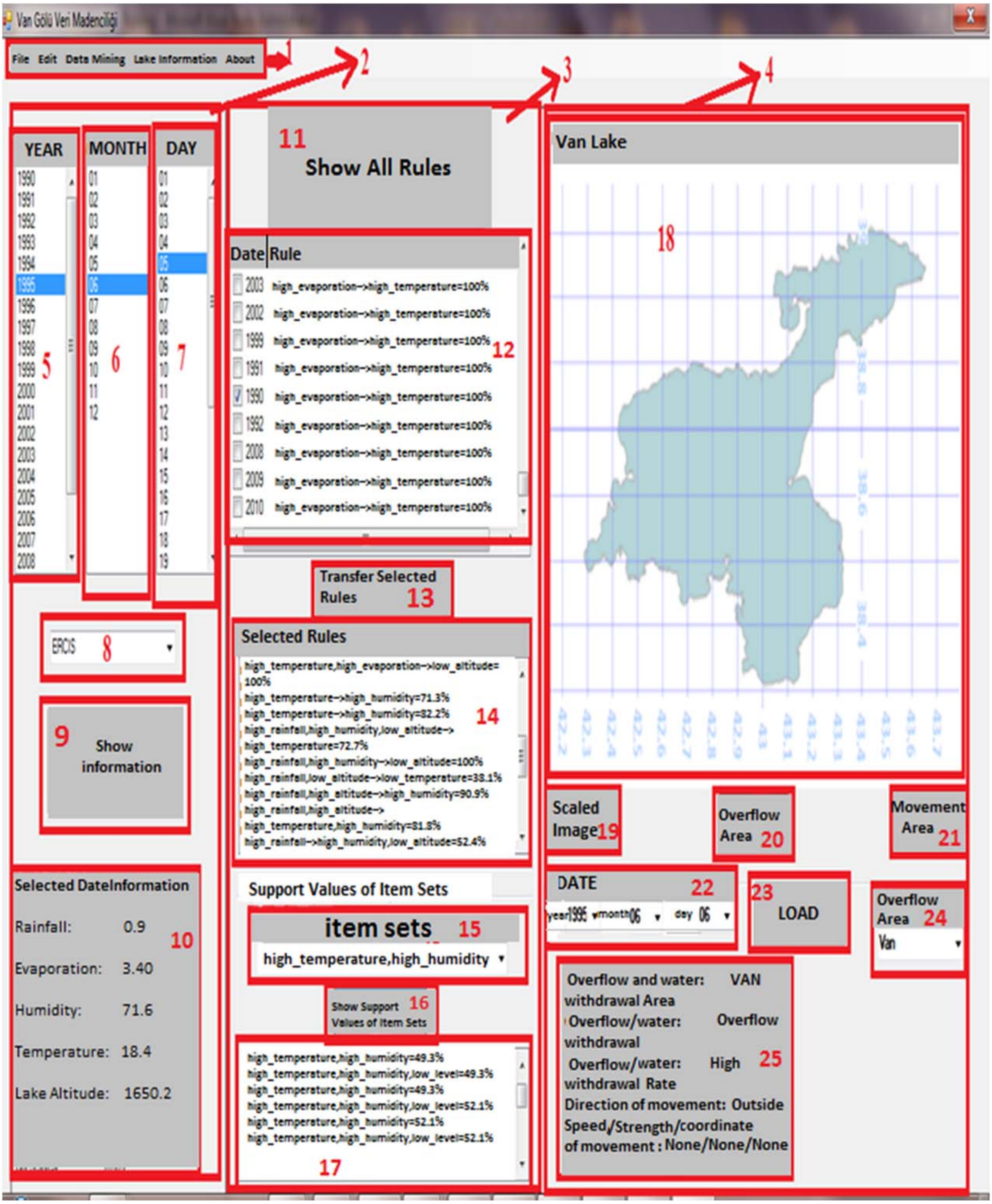

Fig. 4 Application of spatio-temporal data mining with association rules

Part 4 gives us spatial information about Lake Van and 28 overflow-underflow areas. This part also gives detailed overflow-underflow information for a selected region. (Lake Van itself, or any of the 28 overflow-underflow areas) By pressing buttons located in Parts 19, 20 and 21; we see the scaled-map of Lake Van, overflow-underflow areas and directions of overflow and underflow events at 28 mentioned regions at Part 18, respectively. Part 25 gives us whether an overflow-underflow-none occurred at a specific date for the selected overflow-underflow area. It also gives us the magnitude and the direction of the overflow-underflow event. The specific date and the name of the overflow-underflow area are selected in Parts 22 and 24, respectively. The "Show" button located in Part 23 is used to show the detailed overflow-underflow information in Part 25.

To summarize, the association rules in the form of A→ B tell us that the presence of the items in A (also called as "body") are likely to result in the presence of items in B (also called as "head"). In our study, the items are lake altitude, amount of rainfall, evaporation on the surface of the lake, humidity and temperature parameters. The values of these parameters are fuzzified as shown in Table I. Each parameter takes 2 fuzzy values: low or high. Therefore, we have 5 x 2 = 10 items to be used in association rules mining process.

TABLE I
FUZZIFIED VALUES OF METEOROLOGICAL PARAMETERS

Parameter	Values
lake altitude	low_ altitude, high_ altitude
amount of rainfall	low_rainfall, high_rainfall
evaporation on the surface of the lake	low_evaporation, high_evaporation
Humidity	low_humidity, high_humidity
Temperature	low_temperature, high_temperature

TABLE II
SOME ASSOCIATION RULES (ALONG WITH CONFIDENCE VALUES) WHOSE HEAD PART INCLUDES LAKE ALTITUDE PARAMETER

Rules	Explanations
high_humidity→ low_ altitude (100%)	Existence of high humidity always leads to low lake altitude
high_temperature, high_humidity→ low_ altitude (67,1%)	Existence of high temperature and high humidity generally leads to low lake altitude
high_temperature, high_evaporation→ low_ altitude (100%)	Existence of high temperature and high evaporation always leads to low lake altitude
high_rainfall→ high_humidity, low_ altitude (52,4%)	Existence of high rainfall and high humidity generally leads to low lake altitude
high_evaporation, high_humidity→ high_temperature, low_ altitude (100%)	Existence of high evaporation and high humidity always leads to high temperature and low lake altitude
high_evaporation→ low_ altitude (100%)	Existence of high evaporation always leads to low lake altitude

If we analyze the induced association rules, the ones whose part B is about the lake altitude may lead us to predict about possible overflow and underflow events. That is, if part B of any association rule says that lake altitude will be low when items in part A are present; we can conclude about a possible underflow. Conversely, if part B of any association rule says that lake altitude will be high when items in part A are present; we can conclude about a possible overflow. We can then warn responsible public and private institutions to take precautions when situations in part A are present and part B is about the altitude of the lake. Table II shows some of such rules.

Part 4 of the software is just for validation purposes. In Part 4, we can obtain the detailed overflow-underflow information for a specific region at a given date. This information is validated by the values of meteorological parameters given in Part 1 for the same date and by the suitable induced association rules in Part 3. During the validation phase, we select the meteorological parameter values measured by the station which is nearest to the selected overflow-underflow area.

III. CONCLUSION

Lake Van, the largest lake of Turkey, is a closed basin. This feature causes the volume of the lake to increase or decrease as a result of change in water amount it holds. In this study, evaporation, humidity, lake altitude, amount of rainfall and temperature parameters recorded in Lake Van region throughout the years were used by the Apriori algorithm and a spatio-temporal data mining application was developed to identify overflows and newly-formed soil regions (underflows) occurring in the coastal parts of Lake Van. Identifying possible reasons of overflows and underflows may be used to alert the experts to take precautions and make the necessary investments.

REFERENCES

[1] Özçakır F.C. and Çamurcu A.Y., 2007. Birliktelik Kuralı İçin Bir Veri Madenciliği Yazılım Tasarımı ve Uygulaması, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 12, 21-37.
[2] Han, J. and Kamber, M., 2006. Data Mining Concepts and Techniques. Morgan Kauffmann Publishers Inc, 865, San Francisco.
[3] Fayyad U.M., Piatetsky-Shapiro G. and Smyth P., Data Mining and Knowledge Discovery in Databases: An overview, *Communications of ACM*, 39:11, November 1996.
[4] Piatetsky-Shapiro G. and Frawley W., 1991. Knowledge Discovery in Databases, AAAI/MIT Press.
[5] Agrawal, R. and Srikant, R., 1994. Fast Algorithms for Mining Association Rules, *In Proceedings of the 20th International Conference on Very Large Databases*, 487-489, Santiago, Chile.
[6] Yıldız M.Z. and Deniz O., 2005. Kapalı Havza Göllerinde Seviye Değişimlerinin Kıyı Yerleşimlerine Etkisi: Van Gölü Örneği. *Fırat Üniversitesi Sosyal Bilimler Dergisi*, 15(1).