A long-term runoff forecast model based on association rules of data mining

FUQIANG WANG & SHIGUO XU

Institute of Environmental and Water Resources, Department of Civil Engineering, Dalian University of Technology, Dalian 116024, China

sgxu@dlut.edu.cn

Abstract Association rule is an important method for data mining. There are a lot of hydrological and forecasting data in the region area of long-term runoff forecasts. It is important to fully analyse and mine these data via various intelligent algorithms to formulate hydrological forecast for a precise forecast. Considering the characteristics of hydrological forecasting, the association rule mining method is applied to the long-term runoff forecast. The hydrological and meteorological data from 1956 to 2005 were selected to constitute the Jiangqiao station runoff forecast database at Nenjiang River. According to the min-support and min-confidence, the data was pre-treated before extracting association rules by the standards to find the strong association rules. In the practical example of Jiangqiao station, three strong association rules were mined and these rules reveal the effects of the North Pacific sea surface temperature (SST) on the flood season runoff at Jiangqiao hydrological station. The qualified rate of the model is 80%, and the results show that the model is highly effective for flood prediction of Jiangqiao station in the flood season. Furthermore, the association rule mining may be used as an effective tool for the long-term hydrological forecast.

Key words association rules; data mining; sea surface temperature (SST); long-term runoff forecast

INTRODUCTION

With the development of long-term hydrological forecasting and the increasing use of very large databases and data warehouses, the number of species and amount of hydrometeorological data has been increasing sharply. Revealing valuable knowledge hidden in massive data becomes more critical for decision making. When more data is collected and accumulated, extensive data analysis would be easier with effective and efficient data mining methods. How to store, manage and analyse these data has become an important problem in the research of long-term hydrological forecasts. How to extract the useful information and make long-term hydrological forecasts are more problems. The conventional long-term hydrological forecast pattern cannot deal with the numerous data. Furthermore, we expect that data is gathered and analysed by using the computer from the viewpoint of information management. Therefore, extraction of available implicit information to aid decision making has become a new and challenging task.

Data mining is a new research area that aims to extract implicit, previously unknown and potentially useful information from databases (Chen *et al.*, 1996; Abascal *et al.*, 2006; Karasozen *et al.*, 2006; Perner, 2006; Lavrac *et al.*, 2007). Many approaches have been proposed to extract data. Mining association rules are one of the most important. The problem of finding association rules was first introduced by Agrawal *et al.* (1993). This problem aims at discovering unknown relationships, providing results that can be the basis of forecast and decision making (Zaki, 2004). For example, supermarkets or catalogue companies collect sales data from sale orders. These orders usually consist of the sale date, the items in the transaction, and customer-ID. Through getting and analysing of association rules of these orders, a lot of valuable information such as the buying patterns of consumers can be inferred (Burdick *et al.*, 2005). Finding frequent item-sets is the most fundamental and essential problem in finding association rules, from which association rules can be generated directly (McGarry, 2005).

Considering the characteristics of hydrological forecast, the association rules mining method is applied to the long-term runoff forecast. A long-term runoff forecast model has been set up based on the association rules mining method, and it indicated that the method might be an effective tool for long-term hydrological forecasting by taking the flood season runoff forecast of Jiangqiao station as an example.

The rest of the paper is organized as follows: related works are reviewed and the contribution of this paper is discussed; the application of association rules techniques in long-term runoff forecast is described by taking the flood season runoff prediction of Jiangqiao station as an example; and conclusions.

RELATED WORKS

Conception of association rules

Association rules were introduced in Agrawal *et al.* (1993) as a method to find relationships among the attributes in a database. Using these techniques interesting qualitative information with which we can make later decisions, can be obtained. In general terms, an association rule is a relationship between attributes in the way $C_1 \Rightarrow C_2$, where C_1 and C_2 are pair conjunctions (attribute-value) in the way $A = v$ if it is a discrete attribute or $A \in [v_1, v_2]$ if the attribute is continuous or numeric. Generally, the antecedent is formed by a conjunction of pairs, while the consequence is usually a unique attribute-value pair.

There are a large number of rules of this kind in most databases, so it is essential to define some measures that allow us to only filter the most significant ones. The most used measures to define the interest of the rules were described in Agrawal *et al.* (1993):

- *Support*. It is a statistical measure that indicates the ratio of the population that satisfies both the antecedent and the consequent of the rule. A rule R: $C_1 \Rightarrow C_2$ has a support s, if s% of the records of the database contain C_1 and C_2.
- *Confidence*. This measure indicates the relative frequency of the rule, that is, the frequency with which the consequent is fulfilled when it also fulfilled the antecedent. A rule R: $C_1 \Rightarrow C_2$ has a confidence c, if the c% of the records of the database that contain C_1 also contain C_2.

The goal of the techniques that search for association rules is to extract only those that exceed *min-support* and *min-confidence* that are defined by the user. The basic data mining process may be summarized as follows: (a) find out all frequent itemsets and their support counts. A frequent itemset is a set of items which are contained in a sufficiently large number of transactions, with respect to a minimum support. (b) From the set of frequent itemsets found, find out all the association rules that have a confidence value exceeding a minimum confidence. Some of these algorithms can be seen (Agrawal *et al.*, 1994; Manila *et al.*, 1994; Park *et al.*, 1995; Savarese *et al.*, 1995).

Process of study

Our goal is to find association rules in a runoff forecast database, which have a big effect on the runoff. This work can be summarized in four phases: (a) The factors that exceed critical correlation coefficients are selected based on the correlative analysis between SST and flood season runoff of Jiangqiao station. (b) Data of "key months" SST and flood season runoff of Jiangqiao station is discretized to build the prediction databases. (c) The strong association rules which accord with the min-support and min-confidence can be found by use of Apriori algorithm (Aflori & Craus, 2007; Rodtook & Makhanov, 2007). (d) The long-term runoff forecast model is built by explaining the strong association rules.

PRACTICAL IMPLEMENTATION

Research area and data

Songhua River is one of the seven major Chinese rivers. Jiangqiao station on the Nenjiang River, which is a main tributary of Songhua River (Fig. 1). The annual mean precipitation is 400–600 mm in the Nenjiang River basin. Precipitation is concentrated in the summer (June to August) and

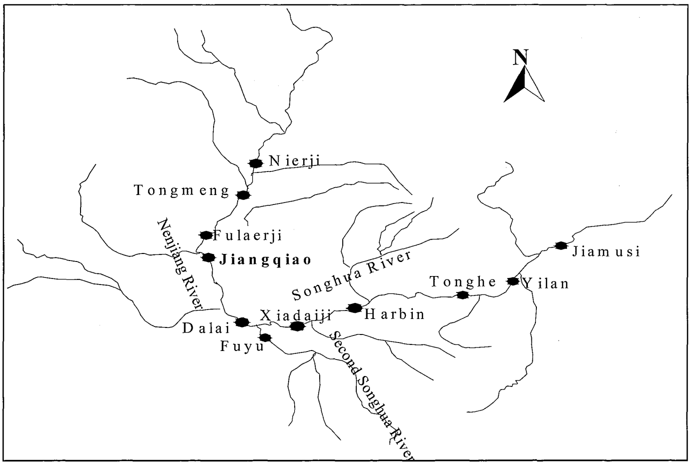

Fig. 1 Location of Jiangqiao station at Nenjiang River.

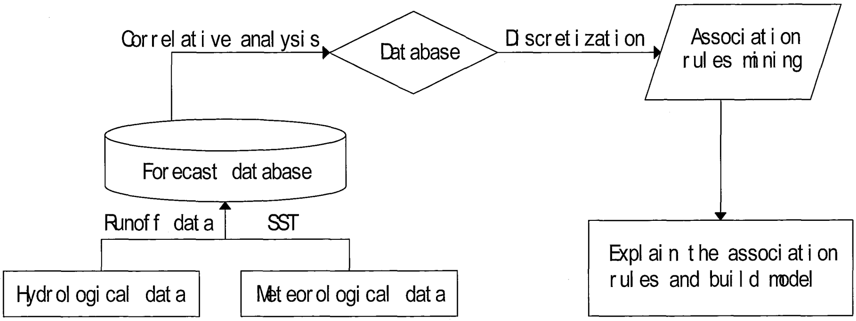

Fig. 2 The process of association rules mining

accounts for 80% of the precipitation in the whole year. In this situation it is important to prevent floods in the flood season.

Fifty years (1956–2005) of continuous monthly runoff data from Jiangqiao hydrological station are used in the study. Monthly SST data from the north pacific region, practically located in the area that spans from 10°S, 50°N and 120°E, 80°W (286 dots, 5° × 5°) are used. The data ranges from 1951 to 2005.

Process of association rules mining

Firstly, the physical factors whose correlation coefficients beyond critical correlation coefficients can be selected based on the correlative analysis between SST and flood season runoff of Jiangqiao station. After that, data of "key months" SST and flood season runoff of Jiangqiao station is discretized to build the prediction database. Thirdly, the strong association rules which accord with the min-support and min-confidence can be found by use of the *a priori* algorithm. Finally, the long-term runoff forecast model is built by explaining the strong association rules (Fig. 2).

Database

We should set up the long-term runoff forecast database first in order to extract association rules. Considering the delay effects of equatorial SST on the climate of research area (Liu *et al.*, 2001), the physical factors whose correlation coefficients are beyond critical correlation coefficients (0.4) can be selected for runoff prediction. We can find that SST of April, May, June and October have a strong effect on the flood season (June to September) runoff of Jiangqiao station via correlative analysis. April, May, June and October are seen as "key months" and their data is selected to set up the long-term runoff forecast database in the study.

Data pre-treatment

The data should be pretreated before extracting association rules. Runoff data for the period 1956–2000 can be classified based on the standard for hydrological information and hydrological forecasting (SL250-2000, 2000) (see Table 1).

Table 1 The classification standard for long-term qualitative flood forecast.

Classification	Low-runoff (1)	Lower-runoff (2)	Mean-runoff (3)	Higher-runoff (4)	High-runoff (5)
Factor Anomaly (FA) (%)	FA < –20	–20 ≤ FA < –10	–10 ≤ FA ≤ 10	10 < FA ≤ 20	FA > 20

The classification standard for SST data can be seen in Table 2. The data of the north pacific sea surface temperature is classified as three ranks.

Table 2 The classification standard for SST data.

Classification	Low-SST (–1)	Mean-SST (0)	High-SST (1)
Factor Anomaly(FA) (°C)	FA <–0.5	–0.5 ≤ FA ≤ 0.5	FA > 0.5

The database is set up based on the discrete data, as can be seen in Table 3.

Table 3 The database of long-term flood forecast.

Number	Year	Physical factors SST(April–June)	SST(October)	Jiangqiao Runoff
1	1956	–1	–1	5
2	1957	–1	–1	5
3	1958	1	1	3
4	1959	1	0	1
5	1960	0	0	5
—	—	—	—	—
45	2000	–1	–1	1

Table 4 Association rules.

Rules	Support	Confidence
$SST(4-6)=1$ and $SST(10)=1 \Rightarrow runoff=5$	0.156	0.636
$SST(4-6)=-1$ and $SST(10)=-1 \Rightarrow runoff=1$	0.178	0.727
$SST(4-6)=-1 \Rightarrow runoff=1$	0.222	0.667

Association rules mining

The goal of the techniques that search for association rules is to extract only those that exceed *min-support* and *min-confidence* that are defined by the user. *A priori* algorithm is used to extract association rules in the long-term runoff forecast database in the study. The results that are discovered by using *a priori* algorithms are shown in Table 4.

Experimental results

- *Rule1: $SST(4-6) = 1$ and $SST(10) = 1 \Rightarrow runoff = 5$ (15.6% support, 63.6% confidence)*
  This rule can be expressed as $SST(4-6) = 1$ *and* $SST(10) = 1 \Rightarrow runoff = 5$ *(15.6% support, 63.6% confidence)*, where 63.6% is the confidence level of the rule and 15.6% is the support level of the rule, indicating how frequently it appears simultaneously as both SSTA(4 – 6) > 0.5°C $\wedge$ SSTA(10) > 0.5°C in the previous year and high runoff of Jiangqiao station in the next year.
- *Rule2: $SST(4-6) = -1$ and $SST(10) = -1 \Rightarrow runoff = 1$ (17.8% support, 72.7% confidence)*
  Rule2 shows that sea surface temperature in April–June and October are all at a low level (<0.5°C) in the previous year, the flood season runoff of Jiangqiao station is also at a low level (runoff anomaly<20%) in the next year, where the support level is 17.8% and the confidence level is 72.7%.
- *Rule3: $SST(4-6) = -1 \Rightarrow runoff = 1$ (22.2% support, 66.7% confidence)*

The rule reveals that the sea surface temperature of April to June is at a low level (<0.5°) in the previous year; the flood season runoff of Jiangqiao station is also at a low level (runoff anomaly >20%) in the next year, where 22.2% of support level and 66.7% of confidence level.

The results of the experiment indicate that when the sea surface temperature of April to June and October are all at a high level (>0.5°) or a low level (<0.5°) in the previous year, the flood season runoff of Jiangqiao station is also at a high-runoff level (runoff anomaly□20%) or a low-runoff level (runoff anomaly <20%) in the next year where there is higher support and confidence. At the same time, the changes of sea surface temperature of April to June have a more evident effect on the flood season runoff of Jiangqiao station in the next year. So more attention should be paid to the changes of SST of "key months" when we make a flood prediction. The other changes of SST do not have a significant effect on the flood season runoff of Jiangqiao station because the other rules cannot satisfy with *min-support* and *min-confidence*.

Model test

Five years (2001–2005) of continuous runoff data from Jiangqiao station are used to test the model in the study. The results can be seen in Table 5. From Table 5, we can see that the prediction results of 2001, 2002, 2003 and 2004 agree well with the measurement. The results of tests show that the qualified rate of the model comes to 80%, and the model is highly effective for the flood prediction of Jiangqiao station in the flood season. Furthermore, association rules mining may be an effective tool for long-term hydrological forecast.

Table 5 The results of the model test.

Year	Physical factors		Measurement	Prediction	Evaluation
	SST(April–June)	SST(October)			
2001	–1	–1	1	1	√
2002	–1	0	1	1	√
2003	1	1	5	5	√
2004	–1	0	1	1	√
2005	1	0	1	–	×

CONCLUSIONS AND FUTURE WORK

In this paper, the association rules mining method is used to discover the unknown, valuable and operable information in physical factors for long-term runoff prediction.

The study has two major contributions. The first is that some interesting rules, which have a strong effect on the changes of flood season runoff of Jiangqiao station, have been discovered.

These rules are useful for the flood prediction of Jiangqiao station in the flood season. The second contribution is that a new prediction method is introduced and identified for long-term hydrological forecasting. How to fully analyse and mine those data via various intelligent algorithms to formulate hydrological forecast and reservoir operation algorithm accordingly for precise forecasts and rational operation is revealed in this paper. Meanwhile, the paper demonstrates that association rules mining is an effective data mining analysis tool in long-term hydrological forecasts.

However, the paper has some possible extensions. How to express the rules in higher level concepts, etc. is very important for practical purposes. Without generalization, the generated rules may have too much detail and not be fit for decision makers. Therefore, by including the concept hierarchies we may produce rules that are more abstract and concrete. Another possible extension is to prune less interesting rules that are trivial or implied by other rules. Without removing the uninteresting rules, we may be overwhelmed by enormous rules. Besides, we may try to set up a database that includes many kinds of physical factors, such as SST, atmosphere circulation characteristics volumes, the height of 500 hpa and 100 hpa isobaric surface, etc. This will help us to make a more precise forecast and rational operation.

Acknowledgements The authors thank Mr Fan Chuiren (Changchun Natural Calamity Forecast and Research Service Center, Changchun, China) for offering SST and runoff data in this study. This study was partially supported by the National Natural Science Foundation of China (no. 50139020).

REFERENCES

Abascal, E., Garcia, L. I. & Mallor, F. (2006) Data mining in a bicriteria clustering problem. *European J. Operational Res.* **173**(3), 705–716.

Aflori, C. & Craus, M. (2007) Grid implementation of the *a priori* algorithm. *Adv. Engng Software* **38**(5), 295–300.

Agrawal, R., Imielinski, T. & Swami, A. (1993) Mining association rules between sets of items in large databases. In: *Proc. SIGMOD International Conference on Management of Data.* 207–216. Washington, DC, USA.

Agrawal, R. & Srikant, R. (1994) Fast algorithms for mining association rules. In: *Proc. Of the VLDB Conference.* 487–489. Santiago, Chile.

Burdick, D., Calimlim, M., Flannick, J., Gehrke, J. & Yiu, T. M. (2005) A maximal frequent itemset algorithm. *IEEE Trans Knowledge Data Engng* **17**(11), 1490–1504.

Chen, M.S., Han, J. & Yu, P. S. (1996) Data mining: an overview from a database perspective. *IEEE Trans Knowledge Data Engng* **8**, 866–883.

Karasozen, B., Rubinov, A. & Weber, G. (2006) Optimization in data mining. *European J. Operational Res.* **173**(3), 701–704.

Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. & Kobler, A. (2007) Data mining and visualization for decision support and modeling of public health-care resources. *J. Biomedical Informatics* **40**(4), 438–447.

Liu, S. & Wang, N. (2001) The impacts of antecedent ENSO event on air temperature over northeast China in summer. *J. Tropical Meteorol.* **17**(3), 314–315.

Manila, H., Toivonen, H. & Verkamo, A. I. (1994) Efficient algorithms for discovering association rules. In: *KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, 181–192. Seatle, Washington, USA.

McGarry, K. (2005) A survey of interestingness measures for knowledge discovery. *Knowledge Engng Rev.* **20**(1), 39–61.

Park, J. S., Chen, M. S. & Yu. P.S. (1995) An effective hash based algorithm for mining association rules. In: *Proc. ACM SIGMOD Int. Conf. on Management of Data.* 175–186. San Jose, California, USA.

Perner, P. (2006) Recent advances in data mining. *Engng Appl. Artificial Intelligence* **19**(4), 361–362.

Rodtook,A. & Makhanov, S. S. (2007) A filter bank method to construct rotationally invariant moments for pattern recognition. *Pattern Recognition Letters* **28**(12), 1492–1500.

Savarese, A., Omiecinski, E. & Navathe, S. (1995) An efficient algorithm for mining association rules in large databases. In: *Proc. of the VLDB Conference.* 432–444. Zurich, Switzerland.

SL250-2000, (2000) Standard for hydrological information and hydrological forecasting. Ministry of Water Resources, China.

Wu, F., Chiang, S. & Lin, J. (2007) A new approach to mine frequent patterns using item-transformation methods. *Information Systems* **32**(7), 1056–1072.

Zaki, M. J. (2004) Mining non-redundant association rules. *Data Mining & Knowledge Discovery* **9**(3), 223–248.