HINTS DRIVEN KNOWLEDGE DISCOVERY AND MANAGEMENT

Hiroshi Sugimura and Kazunori Matsumoto
Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology
1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN

ABSTRACT

This paper is concerned with an information system that carries out datamining and management of discovered knowledge. This system is effectively applicable to many industrial areas, such as finance, process control, and so on. In these areas, there are various kinds of heuristic quasi-knowledge, which are not justified and their usefulness remain ambiguous. We show they are usable as suggestive hints of further knowledge discovery. For this purpose, we propose a new datamining technique that uses decision tress and hints. We further propose a knowledge evolution mechanism, by which we can improve the reliability of the hints and knowledge. The knowledge evolution process includes two main tasks: evaluation and rediscovery with improved hints. The essential idea assumes that better knowledge comes from a better hint, we therefore propose a hint improvement method based on the genetic algorithm. The algorithm produces the improved hints by next generation of applying the standard GA operations, and also directly evolutes the discovered knowledge.

KEYWORDS

Knowledge acquisition, data mining, decision tree, time-series data

1. INTRODUCTION

In accordance with the varieties of data and knowledge, various types of information systems are used in industrial areas. In particular, time-series data is the most commonly appeared in so many practical areas such as stock prices [1], exchange rates, scientific data, sensing data in industrial plants, and so on. Although there exists advanced mathematical theories [2] dealing with these data, artificial intelligence approaches are also important. We have strong requirements for information systems that behave intelligently by using experimental and heuristic knowledge. Developments of that type of systems are now actively going on. Among many technologies in artificial intelligence, datamining [3, 4, 6] is one of the most expected practices, which aims at an extraction or discovery of useful and unknown knowledge from a large amount of data.

We point out here that in practical areas there are various kinds of heuristic quasi-knowledge. They are relating to patterns of data, but are not justified and their usefulness remains ambiguous. In order to make the best use of such quasi-knowledge, we propose a new datamining technique that uses decision tress [7, 11] and quasi-knowledge. The value of quasi-knowledge is relative to a situation, which includes a purpose of quasi-knowledge, user's expectation, etc. A direct application of decision tree learning deals with *n* values on *n* time points as *n* attribute values. This approach focuses only on the separated properties of time-series data, thus we ignore continuous behaviors of them. To avoid this problem we need to investigate patterns appearing in time-series. Several studies point out that a brute force exploration often identifies meaningless simple behaviors like a cosine curve. A useful and non-trivial pattern is hard to identify automatically. Then we use human experience, which is expressed in quasi-knowledge, as hints of useful and non-trivial patterns. They become features on attributes in the learning. In another characteristic of time-series data, we need a method to manage ambiguity in data. We are necessary to identify a set of data having similar shapes, different length, different values on some points, etc. For this propose we use a technology that is developed in image processing. This short paper also shows an outline of the entire system and a methodology. Figure 1 shows a workflow of the system.

2. OUTLINE OF SYSTEM

In Figure 1, the arrows show main flows in the datamining process. *DB Miner* discovers knowledge from *Raw DB*, and stores it in *Mined KB*. Quasi knowledge is given to the system through *GUI* and is stored in *Quasi KB* with the reliability grade is set to the lowest value. *Template DB* stores templates that control the datamining algorithm in *DB Miner*. By using knowledge in *Quasi KB* and *Mined KB*, *Predictor* makes a prediction about a future situation based on data in *Raw DB* and current real-time data obtained through *I/F*. The prediction is evaluated by a human expert. According to the evaluation result, *Selector* adjusts the reliability value for each part of knowledge.

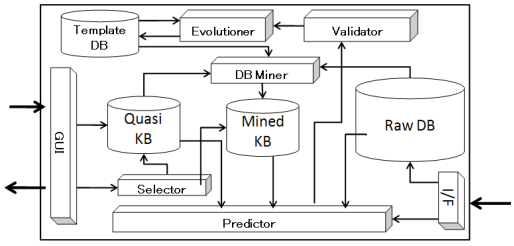

Figure 1. An Outline of System

3. USE OF QUASI KNOWLEDGE

In datamining, the usefulness is usually decided based on numerical values calculated from predefined evaluation rules. The confidence and support [8, 9] are widely used for this purpose. The templates in our system are prepared to distinguish desirable rules from bad rules. Knowledge that is determined bad by the templates are discarded. Similar idea can be found in [5, 10, 12], but they applied to simple discrete valued transactional databases. Our method is extended to deal with time-series data, which are finite sequences of continuous values. We thus develop a new method that includes an expression language of time-series template, and an algorithm that finds a portion of knowledge which matches to a specified template. In contrast to the simple discrete case, finding a match among time-series data becomes a difficult problem. The basic idea is a use of dynamic time-warping, and a calculation of similarity scores depending on several parameters. The parameters must be determined considering the characteristics of data and knowledge in the domain. Methods in reinforcement learning are applicable to adjusting the parameters.

Simple examples of quasi-knowledge are shown, p_1, p_2, and p_3, in Figure 2(a), which state specific behaviors of data. These are given to *DB Miner*, and are used as 'focus points' in the mining algorithm. In the case of association rule mining, rule candidates which contain similar patterns to these are examined first. As a result, we would obtain knowledge as shown in Figure 2(b). Starting with quasi-knowledge, we synthesize more useful knowledge. By integrating this method with the templates, we more increase the performance.

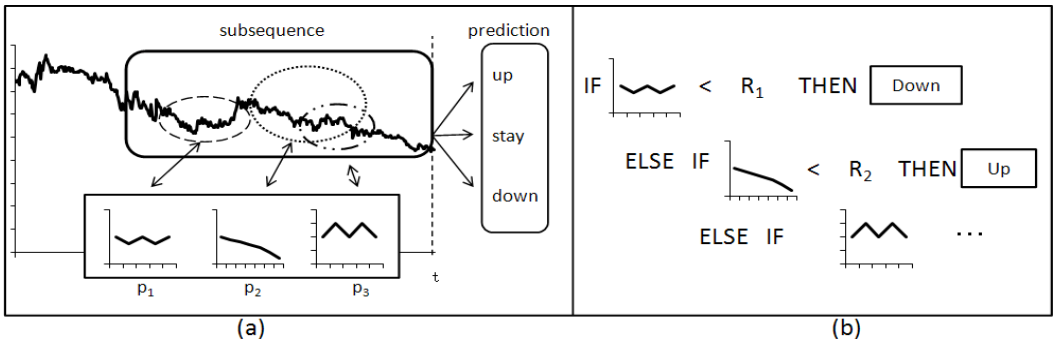

Figure 2. Example of Quasi-Knowledge and Extracted Rules

4. KNOWLEDGE EXPRESSION

Time-series data is a sequence of real numbers $V = v_1,...,v_i,...,v_n$, where n is the length of this data. This type of data, such as temperatures, stock prices, sensing values, and so on, occurs widely in various practical fields.

We apply the dynamic time warping [2], DTW for short. For two, sequences having different lengths i, and j, the similarity degree $g(i, j)$ is defined in the following, where $i-1$ and $j-1$ is the sequences removing the last values.

$$g(i, j) = \min\{g(i, j-1) + q, g(i-1, j) + r, g(i-1, j-1) + s\}$$

$$s = |a - b|$$

In the equation, q and r represent the cost of shortening and expanding the sequences along the time axis, and s is the distance cost of values. More similar pair has a smaller value, and it becomes zero if they are exactly the same. For a given threshold value, we regard they are equal. The value is called a maximum allowance.

5. QUASI-KNOWLEDGE EVOLUTION

As we described above, a set of knowledge works as hints of knowledge discovery. We assume that better knowledge is brought by using better hints, we therefore propose an evolution mechanism of hints. The essential idea is based on a use of the genetic algorithm, GA for short, Figure 3 shows an outline of the evolution process including several tasks, and Figure 4 shows a rough sketch of the GA task. As we described above a set of quasi-knowledge becomes features in decision tree learning. The evolution task applies three basic GA operations, the selection, crossover and mutation, then produces the next generation quasi-knowledge. They again become new features in the learning. We repeat this task until the decision tree reaches a prefixed performance. Another new proposal is a direct evolution of the decision tree. This outline is shown in Figure 5. Table 1 shows an experimental result.

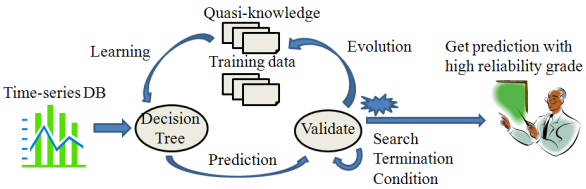

Figure 3. An outline of evolution process

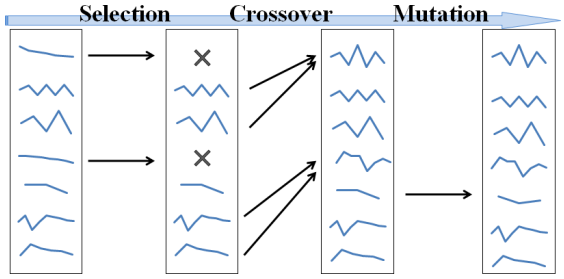

Figure 4. An outline of GA task

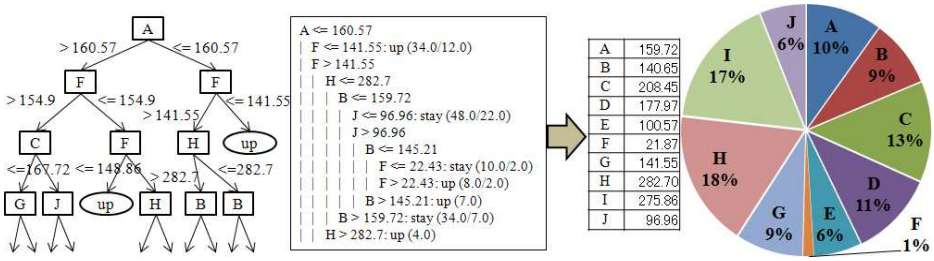

Figure 5. A direct evolution of the decision tree

Table 1. An experimental result

stock	before	after	improve
Electric Company 1	73.973	79.452	5.479
Electric Company 2	74.429	79.452	5.023
Electric Company 3	67.884	73.973	6.089
Electric Company 4	69.863	76.865	7.002
Electric Company 5	63.166	69.711	6.545
Electric Company 6	82.801	82.801	0.000
Electric Company 7	73.059	78.387	5.328
Electric Company 8	71.799	78.049	6.250
Electric Company 9	60.161	70.323	10.162
Electric Company 10	69.559	79.951	10.392
Electric Company 11	67.835	73.781	5.946
Electric Company 12	61.738	68.140	6.402
Electric Company 13	60.883	68.037	7.154
Electric Company 14	70.167	76.408	6.241
Electric Company 15	75.343	79.452	4.109

6. CONCLUSIONS

This paper shows an outline of the datamining system that uses quasi-knowledge as hints. The quality of extracted knowledge depends on that of given quasi-knowledge. We thus propose a evolution mechanism to improve them to start up the discovery task. We investigate these proposals by using more than ten years of stock market data. The experimental results will be shown in the presentation.

REFERENCES

[1] http://stockcharts.com/

[2] Donald J.Berndt, James Clifford, 1994. Using Dynamic Time Warping to Find Patterns in Time Series. *In Proceedings. of KDD Workshop*, Washington, USA, pp.359-370.

[3] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, 2008. *Time Series Analysis: Forecasting and Cotrol*, John Wiley & Sons Inc, San Francisco, USA.

[4] Ian H. Witten, Eibe Frank, 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, San Francisco, USA.

[5] Jiawei Han, Micheline Kamber, 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, USA.

[6] J. M. W. Tadion, 1996. *Deciphering the Market: Principles of Chart Reading and Trading Stocks, Commodities, and Currencies*. John Wiley & Sons Ltd, San Francisco, USA.

[7] J. Ross Quinlan, 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Fransisco, USA.

[8] M. A. Bramer, 2007. *Principles of Data Mining*. Springer, New York ,USA.

[9] Nils J. Nilsson, 1998. *ARTIFICIAL INTELLIGENCE: A New Synthesis*. Morgan Kaufmann Publishers, San Francisco, USA.

[10] Mark Last, Abraham Kandel, Horst Bunke, 2004. *Data Mining In Time Series Databases*. World Scientific Pub Co Inc, New Jersey, USA.

[11] Ron Kohavi, 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid, *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, California, USA, pp.202-207.

[12] Trevor Hastie, Robert Tibshirani, J H Friedman, 2001. *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York, USA.

[13] Yasushi TANAKA, Yoshiaki TAKANO, Hiroshi SUGIMURA, Kazunori MATSUMOTO, 2009. Knowledge Acquisition and Maintenance based on Data Mining with Preferences. *In Prceedings of the Iadis International Conference Information Systems 2009*, Barcelona, Spain, pp.469-472.

[14] Yoshiaki TAKANO, Yasushi TANAKA, Hiroshi SUGIMURA, Kazunori MATSUMOTO, 2009. Employing Empirical Knowledge in Practice. *In Prceedings of the Iadis International Conference Information Systems 2009*, Barcelona, Spain, pp.481-484.