DATAMINING TOOL WITH EXPLORATORY SEARCH AND FEATURE DISCOVERY

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 proposes a system that datamines knowledge to extrapolate future behaviors of time-series data. This system includes two major tools. First one is a search tool which collects necessary information from databases. The standard database query language is extended to deal with typical properties of time-series data. We explain an overview of the language and a search mechanism. Second tool provides an automatic mechanism to discover features over a given set of training data. We can describe data in terms of discovered features. The discovery procedure begins with a set of randomly generated features and successively improved, generation by generation, using the genetic algorithm. By using the well optimized features we build a decision tree that predicts future behaviors. We explain how these two tools are combinatory applied in the entire knowledge discovery process.

KEYWORDS

data mining, machine learning, query, genetic algorithm, time-series data.

1. INTRODUCTION

This paper proposes a system that datamines knowledge to extrapolate future behaviors of time-series data. As shown in Figure 1, the dataminig process consists of several different tasks [1, 4, 9, 10]. In these tasks, necessary data are collected interactively from databases, are preprocessed, and are transformed [3, 7, 11]. For this purpose, we need a powerful database and data manipulation mechanism. In the case of time-series data, the standard database query language SQL is insufficient. Thus we develop an extended query language, which is named Time-series Query Language, TQL for short. Another important purpose in the data transformation task is to describe data in terms of features. In general, discovery of features becomes a crucial problem which depends on human experts knowledge [14]. To solve the problem, we develop a tool to automatize the identification task. In the following chapters, we explain outlines of TQL and the automatized tool.

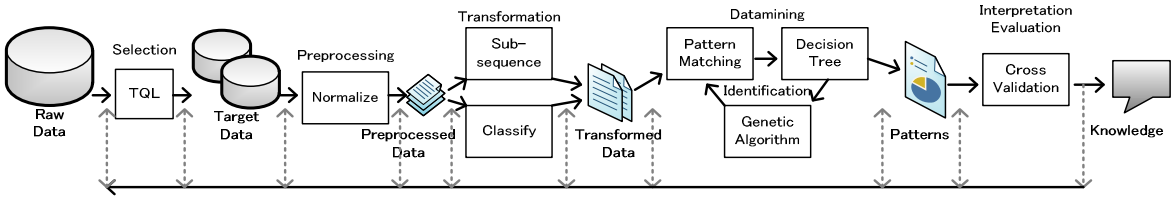

Figure 1. Outline of system

2. TIME-SERIES QUERY LANGUAGE

The time-series query language, TQL, can deal with a rough sketch of pattern given by a handwriting. A query of TQL is constructed by a combination of patterns and the operators which is listed below in Table 1

and Table 2. To maintain the simplicity and the easiness in learning of TQL, these operators are designed similar to the ones of XPath. For example, an orders of occurrences of patterns is specified by using "/". Similarly, regular expressions on patterns are expressed in the standard way. In Figure 2, we show an example of queries and search results.

Table 1. Syntax of TQL of queries

Query ::= Path	Path Op Query
Path ::= Pattern	Pattern "/" Path
Pattern ::= String	wildcard
Op ::= "and"	"or"
Wildcard ::= Refer to Table 2	

Table 2. Wildcard and quantification

.	Matches any single data
*	Zero or more of the preceding element
+	One or more of the preceding element
?	Zero or one of the preceding element
{n,m}	Matches the preceding element at least m and not more than n time

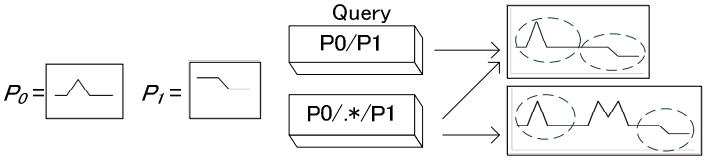

Figure 2. Example of query and searched sequence

3. IDENTIFICATION OF FEATURES AND DECISION TREES

As we show in Algorithm 1, feature discovery task starts with randomly selected candidates. They are improved by the genetic algorithm [12], whose essential idea is shown in Figure 3. The fitness function used in the algorithm is also shown in the procedure 1.

Algorithm: Datamining algorithm based on GA

input: a time-series data T and subsequence size ws
output: the set of all patterns in elite population
begin
$population_1$ = InitializePopulation();
seg = SlidingWindow(T, ws);
c_seg = Classify(seg); // future of seg
elite = 1;
for g = 1 **to** termination condition // ex. g ≤ g_max
  $fitness_g$ = eval all fitness($population_g$, c_seg);
  $population_{g+1}$ = GA_selection($population_g$, $fitness_g$);
  $population_{g+1}$ = GA_crossover($population_{g+1}$);
  $population_{g+1}$ = GA_mutation($population_{g+1}$);
  $Tree_{g+1}$ = DecisionTreeLearning($population_{g+1}$);
  if accuracy of $Tree_{g+1}$ ≤ accuracy of $Tree_{g+1}$
  then elite = g+1 **end;**
end;
output($population_{elite}$);
end;

Algorithm 1. Pattern discovery by using genetic algorithm

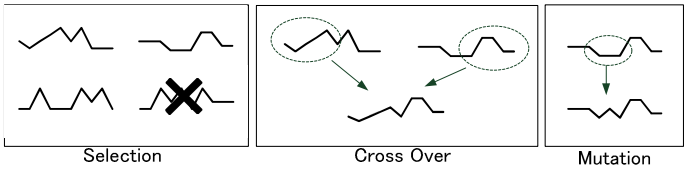

Figure 3. Behavior of genetic operator

Once the features are identified, we apply the decision tree learning algorithm [6, 8] to extract knowledge saying future behaviors. The result of an experiment is shown in Table 3 and Figure 4. This experiment uses stock price data in the Tokyo market in JAPAN.

Procedure: fitness

input: $gene_m$ and a set c_seg of classified subsequences
output: fitness of $gene_m$
begin
  foreach c_seg_m in c_seg
    $dissim_m$ = DTW($gene_n$, c_seg_m);
  end;
  sort c_sig by dissim;
  fitness = 0;
  foreach x is $split\_point_x$ that is dissim

$$infor_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times infor(T_i)\ ;$$

$$infor(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} \times \log_2\left(\frac{freq(C_j, T)}{|T|}\right)$$

  $gain(X) = info(T) - infor_x(T)$;
  if fitness < $gain(X)$ **then** fitness = $gain(X)$; **end;**
  end;
end;

Procedure 1. Fitness function

Table 3. An experimental result

	5 genes		10 genes		20 genes	
	Accuracy	Feature	Accuracy	Feature	Accuracy	Feature
Stock01	52.14	5	57.45	10	65.07	14
Stock02	51.30	5	56.38	8	60.95	13
Stock03	50.23	4	52.77	9	62.58	14
Stock04	51.86	5	60.27	9	62.58	14
Stock05	50.11	5	53.27	9	63.49	17
Stock06	53.50	5	59.37	9	62.36	14
Stock07	52.88	5	54.74	7	64.79	16
Stock08	54.60	5	63.25	7	65.51	12
Stock09	60.55	5	60.27	6	70.26	14
Stock10	62.58	5	62.36	5	66.82	13
Average	53.98	4.9	58.01	7.9	64.44	14.1

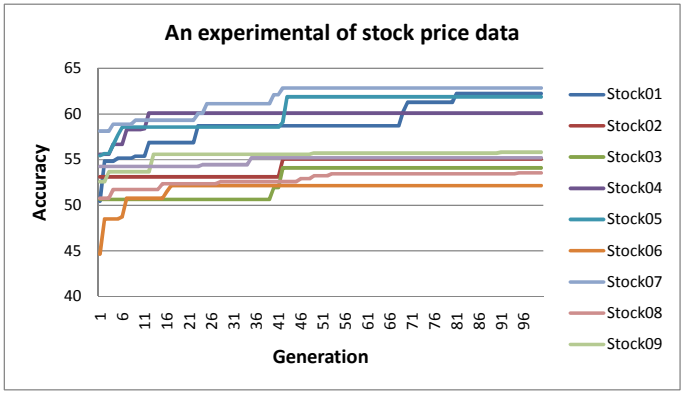

Figure 4. An experimental result

4. CONCLUSION

This paper explains a system that datamines knowledge to extrapolate future behaviors of time-series data. The main points of the system are twofold; use of the extended query language TQL and feature identification based on the genetic algorithm. We explain how these two tools are combinatory applied in the entire knowledge discovery process. Then the effectiveness of the system is demonstrated by using the experimental result.

REFERENCES

[1] Agrawal, R. and Srikant, R. 1995. Mining Sequential Patterns. *Proceedings of the 11th Int. Conf. Data Engineering,* pp.3-14.

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

[5] 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.

[6] 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.

[7] Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. *Proceedings of the 5th Int. Conf. Extending Database Technology,* pp.3-17.

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

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

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

[11] Shigeaki Sakurai, Youichi Kitahara, et. al. 2008. Discovery of Sequential Patterns Coinciding with Analysts' Interests. *Journal of Computers,* Vol 3, No 7, pp.1-8.

[12] John H. Holland, 1975. *Adaptation in Natural and Artificial Systems,* University of Michigan Press.

[13] Agrawal, R., 1993. Database Mining: A Performance Perspective, *IEEE Trans. Knowledge and Data Engineering,* Vol. 5, No. 6, pp. 914-925.

[14] Yasushi Tanaka, Yoshiaki Takano, et. al. 2009. Knowledge Acquisition and Maintenance based on Data Mining with Preferences, *Proceedings of the Iadis International Conference Information Systems 2009*, pp.481-484.