Fifth International Workshop on Computational Intelligence & Applications
IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, November 10, 11 & 12, 2009

Extension of Decision Tree Algorithm for Stream Data Mining Using Real Data

Tatsuya Minegishi†, Masayuki Ise†, Ayahiko Niimi‡, Osamu Konishi‡

†Graduate School of Future University-Hakodate, Systems Information Science
Future University Hakodate. 2, 116 Kameda-nakano-cho, Hakodate, Hokkaido, Japan
Email: {g2109043, g3108001}@fun.ac.jp

‡Future University-Hakodate, School of Systems Information Science
Future University Hakodate. 2, 116 Kameda-nakano-cho, Hakodate, Hokkaido, Japan
Email: {niimi, okonishi}@fun.ac.jp

***Abstract*—Recently, because of increasing amount of data in the society, data stream mining targeting large scale data has attracted attention. The data mining is a technology of discovery new knowledge and patterns from the massive amounts of data, and what the data correspond to data stream is data stream mining. In this paper, we propose the feature selection with online decision tree. At first, we construct online type decision tree to regard credit card transaction data as data stream on data stream mining. At second, we select attributes thought to be important for detection of illegal use. We apply VFDT (Very Fast Decision Tree learner) algorithm to online type decision tree construction.**

I. Introductions

In recently network society, the development of information processing technique enables us to collect and utilize massive amount of data, and the data mining is technology of discovery new knowledge and patterns in those data has been paid attention. But those data are changing from moment to moment, and have become new type large scale data. Record of financial and distributional transactions, telecommunications records and network access logs are typical examples, and those data are called data stream. By data stream, it is that the conditions temporally-changed massive amount of data record are generated, cumulative and consumed are looked on as flow of data (stream). In the real world, the requirement that whenever we need information, we want to elicit from those large scale data stream has been growing.

At first glance data mining seems to be effective, but data stream has following dynamic properties:

(i) massive amount of data are
(ii) coming over high-speed stream
(iii) temporally-changing
(iv) continue to arrive permanently,

and there is a limitation applied data stream to data mining intending static data.

Data mining to efficiently deal in large scale data stream, therefore data stream mining technology has been developed.

In this paper, we conduct data stream mining to deal in real data and propose online type decision tree construction as algorithm of machine learning. For dealing in real data, we conducted verification experiments to use credit card transaction data. We say that those data fit into concepts of data stream from amount of data and contents, and because we use to think that those are most suitable in our experiments.

Additionally we discuss that credit card transaction data have problem of massiveness on analysis. When we detect illegal use on data mining, the massiveness of data has a problem. It's important how to decrease amount of data using analysis to keep detection accuracy. In this paper, we propose attributes selection technique that we construct decision tree from credit card transaction data for preprocessing of data is a process of data mining and analyze using appearing attributes. Therefore we select attributes thought to be important for detection of illegal use from the online type decision tree. And we consider those attributes. We compare the difference of attributes selected from decision tree to attributes using for analysis of existing technique.

The next section presents data mining, data stream mining and credit card transactions. In section Ⅲ we describe proposed algorithm, and in section Ⅳ we describe their verification experiments. From the results, in section Ⅴ and Ⅵ we discuss them results. The last section we conclude our proposed method and discuss our future works.

II. Related Work

A. Offline Type Decision Tree Construction

Decision tree is graphic representation described by tree diagram hierarchized multistage branching process when it multistage and repeatedly executes decision-making or classification of stuff. It started out root node with given data. Then it is pursuing and answering questions concerned attributes of data. It classifies by making value having finally arriving leaf node into class of the data. Offline type decision tree construction is the method of constructing decision tree after taking all examples as input. The most well-known decision

tree algorithm includes C4.5[1]. C4.5 is decision tree construction algorithm developed by J. Ross Quinlan who is Australian researcher. It recurrently splits data while selecting the attribute and the question that classifies data best based on entropy

$$\mathit{info} = -\sum p_i \log_2 p_i \tag{1}$$

where the p_i is the occurrence probability of the i th event for the information gain

$$\begin{aligned} Gain = (&average\ entropy\ before\ splitting \\ &-average\ entropy\ after\ splitting), \end{aligned} \tag{2}$$

and recurrently constructs the decision tree at the same time. However, the error rate is high because it becomes very complicated tree only the structure was caught disregarding meaning content. Therefore, decision tree is pruned to minimize error rate and becomes more simply and easily understandable.

B. Online Type Decision Tree Construction

Offline type decision tree construction in Ⅱ.A is a given fact that it will take all examples as input first. However, this method cannot start constructing without all examples and needs to access randomly to them. Therefore, it cannot apply to data stream.

Decision tree construction method to improve weakness of this offline type is called online type decision tree construction. Representative example includes VFDT(Very Fast Decision Tree learner)[2]. VFDT is decision tree construction algorithm for data stream and adaptively grows the tree without waiting for all examples arrival. It does not store any examples in main memory, requiring only space proportional to the size of the tree and associated sufficient statistics. It can learn by seeing each example at once, and therefore does not require examples from an online stream to ever be stored. VFDT is the same as C4.5 to grow from root node in sequence but every time it takes new data, it sets up leaf node which the data arrive at current tree, and stores arrival data there. Then, the data of often visible type accumulate at leaf nodes, after data to satisfy enough statistical criterion accumulate, it grows the leaf using those data to make a more detail prediction. The statistical criterion to grow leaf moreover includes Hoeffding bound. The data accumulated at leaf is part of all available data, therefore there is a possibility that the leaf has error. However, in consideration of the infinitely-long data stream produced stochastically based on stationary distribution, data sets arriving at each leaves are considered as ideal them in offline case. Consider a real-valued random variable r whose range is R. Suppose we have made n independent observations of this variable, and computed their mean $\overline{r}$. The Hoeffding bound states that, with probability $1-\delta$, the true mean of the variable is at least $\overline{r} - \epsilon$, where

$$\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}} \tag{3}$$

If the difference between the best standard level at one leaf and the next standard level is bigger, then it makes branching from the leaf.

C. Credit Card Transaction Data

In this paper, we use the credit card transaction data as real data. In actual credit card transactions, the data are complex changing and arriving continuously online. The data are following:

(i) arrive around one million transactions per day,
(ii) the speed of less than one second per one transaction,
(iii) arrive around one hundred transactions per one second at the peak time,
(iv) 24 hours a day, every day, continue to arrive permanently.

Therefore, the credit card transaction data can be exactly called a data stream.

However, even if we use data mining for those data, around 2,000 transactions per day which people can accommodate by monitoring are generally. Therefore, we have to detect suspicious transactions data effectively under the rigid conditions of 0.02% detection numbers of the total. In addition, there is an issue people detect extremely low illegal use from massive amount of transaction data because real illegal use is extremely low rate that is from 0.02% to 0.05% to all transaction data.

The data we use in this paper is described one transaction data as CSV format in time order and the data exists as attributes. Credit card transaction data have 124 attributes: 84 are called transaction data include one attribute to discriminate whether the data is illegal use , and the others are called behavior data calculated by user's usage. The file size is about 700 MB per one month. As we said that the illegal use rate is from 0.02% to 0.05% before, this data re-sampled to about 0.5%.

III. Proposal Technique

A. Attribute Selection from Offline Type Decision Tree Construction

When we detect illegal use on data mining, the massiveness of data has a problem. It's important how to decrease amount of data using analysis to keep detection accuracy. In this paper, we propose attributes selection technique that we construct decision tree from credit card transaction data for preprocessing of data is a process of data mining and analyze using appearing attributes.

Additionally, we change number of attributes and type of attributes using analysis to consider some attribute selections. There are following technique as an existing research[3][4].

(i) Constructing decision tree.
  – We construct decision tree using data set of (a) in Ⅳ. Those trees have about the same attribute type and positions of attributes until rank of top five therefore we regard as "stable rank".

(ii) Gathering data that failed in classification of the tree.
  – We collect only data failed classifying until stable rank.

(iii) Constructing decision tree using only gathered data again.
  – We construct decision tree using only those data again

(iv) Adding the attributes for analysis too.
  – When the decision tree constructed using only data failed classifying contains never seen before attributes, we add those attributes to analytic attributes.

We compare the difference of attributes selected from decision tree to attributes using for analysis of existing technique.

B. Attribute Selection from Online Type Decision Tree Construction

In Ⅲ.A we proposed offline type decision tree construction from credit card transaction data, and attribute selection.

Here Ⅲ.B, we propose online type decision tree construction using real data. When a certain level of decision tree constructs, we select attributes in order near from root node of the tree. We regard the real data using Ⅲ.A as stream data like real credit card transaction. We apply VFDT we described in Ⅱ.B to this algorithm. Moreover when we use VFDT, we use VFML(Very Fast Machine Learning)[5] is implementation code of machine learning for data stream, and construct VFDT by it.

In this paper, we definitely don't decide how to select attributes from VFDT. But we have two methods of selection.

The first method, for constructing VFDT in arriving data through stream, we select attributes when the branches grow from certain leaves. This algorithm started from root node. Then it grows the VFDT gradually. When new attributes appear, we select those attributes as analytic attributes.

The other, for constructing VFDT in arriving data through stream, we select appeared attributes when we stop growing VFDT after a certain period of time.

In both cases, it is important to determine when attributes are selected. It is whether decision tree change around stopping growing VFDT is important. These are changed by data arrived through stream. As data for construction are over the long term, there is a possibility that usage tendency between past data and latest data changes significantly. But we got a result that VFDT has no big differences if we construct VFDT using data for about a year. But as data for construction are over the long term, there is a possibility that usage tendency between past data and latest data changes significantly. As a result, it is expected to change VFDT.

IV. Experiments

We constructed online type decision tree from credit card transaction data provided by described technique, and compare accuracy, size and attributes.

A. Construction of Offline Type Decision Tree

We use following data sets in the experiment.

- The number of attributes of data
  - 57 transaction data attributes and 42 behavior data attributes
- Sampling rate of illegal use (three ways)
  (a) 0.02% is actual illegal use rate
  (b) 0.5% is sampling rate of data provided
  (c) 10% is setup in this experiment

Usually provided data have around 120 attribute but we except some attributes are irrelevant for construction decision tree and has low relation for illegal use models. We also use around fifty thousand data.

We construct decision tree using three ways data by J48 algorithm based on C4.5 implemented in data mining tool software called Weka[6].

B. Construction of Online Type Decision Tree

In this experiment, we use the 10% data described as (c) in Ⅳ.A. As we discuss later, this data was the best results on offline type decision tree construction in Ⅳ.A. We construct VFDT by VFML using this data.

V. Experimental Results

In the case of offline type decision tree, (a) became tree split two leaf nodes by root node. (b) became tree that has 101 nodes including 51 leaves.

In Fig.1, we hold showing real attribute names back as confidential information. So we set low resolution consciously. Both (a) and (b) have more than 99 % accuracy but they classify almost all illegal data as normal data. This result shows that the accuracy of decision tree is high, but actually it cannot classify illegal data exactly. Fig.2 has 1,221 nodes including 611 leaves. Fig.2 is set low resolution as well as Fig.1. Also its accuracy is 95.413%. In tis experiment, this tree is able to classify illegal data well because the 10% illegal use rate is higher than another two cases.

In the case of online type decision tree, we constructed it using data (c) for offline type. The accuracy is 92.157% and the size is 91.

VI. Evaluations

We show results of offline and online type decision tree.

TABLE I
Result of Number of Attributes

	All Attributes	**Transaction Attributes**	**Behavior Attributes**
C4.5	55	18	37
VFDT	31	13	18

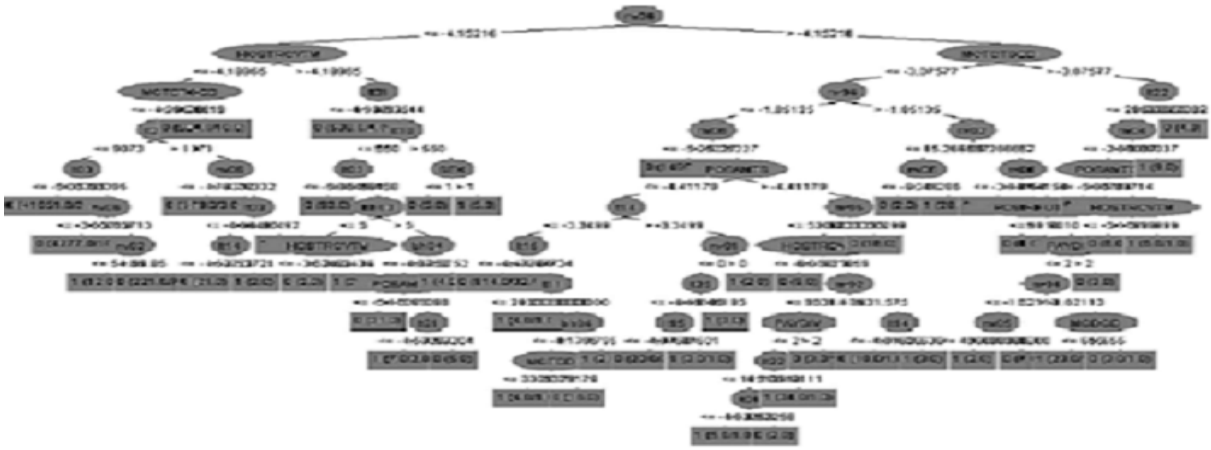

Fig. 1. Decision Tree (b)

TABLE II
RESULTS OF ACCURACY AND SIZE

	Accuracy(%)	**Size**
C4.5	95.413	1,221
VFDT	92.157	91

We constructed both tree random-sampled 10 data without changing illegal use rate. In this Table I and Table II, it is average of 10 trees. Each tree set up the results using 10-folds cross validation, therefore each method constructs actually 100 trees.

In the accuracy of classification, VFDT is lower than C4.5. It is because not construction of decision tree after receiving all examples as input data like C4.5 but growing tree using Hoeffding bound at the point of accumulating adequate data in nodes. The size of VFDT is less than 1/10 of C4.5.

In terms of attributes, 10 offline type decision trees have almost the same feature and positions until 5th depth. Therefore there are no major differences and attributes selected from their trees are important in classification of illegal use. C4.5 has 55 attributes from 99 input attributes using construction. Especially, there are many behavior attributes around the top. However, in spite of having 1,200 nodes, appeared attributes are 55 out of about half of using attributes. As a result, same specific attributes are near root node appear many times.

Also we constructed 10 online type decision trees, but there are 3 types as a root. Moreover, in case of tree having same root, attributes that follow are different from C4.5's attributes and positions. Because those attributes as root node show paying division and limit of amount of usage, they have some features of illegal use. And we asked about this results to experts, and proved that those attributes are appropriate. In online type decision tree, the number of attribute is an average of 31. As well as C4.5, number of attributes got fewer than all data used by construction. Also behavior attributes are more than transaction attributes.

This is the reason that C4.5 and VFDT adopt many behavior attributes as nodes. Therefore behavior attributes affect detection of illegal use.

In this paper, in case of the data that are large-scale and very low illegal use rate for constructing, we don't have a concrete

Fig. 2. Decision Tree (c)

plan for improving accuracy of classification without growing tree structure size. The size of tree of C4.5 is about 1,200 but appeared attributes until stable rank are 14 from 99 input attributes using construction. Also appeared attributes from all nodes of tree are 55.

TABLE III
RESULT OF NUMBER OF ATTRIBUTES

	All Attributes	**Transaction Attributes**	**Behavior Attributes**
Stable rank	14	3	11
All nodes	55	18	37

Table III shows the number of transaction attributes, behavior attributes, and all attributes. For this reason as well as C4.5, appeared attributes are expected to drastically decrease than input attributes using construction in spite of tree structure size for VFDT. Also the detection rate of illegal use keeps more than 80% if analytic attributes decrease to 25%. Therefore it is known that type of attributes is more important rather than the number of attributes.[4]

VII. CONCLUSION AND FUTURE WORK

In this paper, we constructed online type decision tree using credit card transaction data with data stream mining. And we proposed that we select attributes thought to be important for detection of illegal use from the tree. At first, we applied VFDT algorithm to online type decision tree construction. At second, we compared VFDT to C4.5 from focused on accuracy, size and attributes. As a result, the accuracy of VFDT is inferior to C4.5. However, it is simple tree whose size is less than 1/10 to C4.5.

In future work, we are aiming to construct VDFT without using all examples like C4.5. It would appear that we can cut the time for analysis to select attributes at proper time. We need to define rule of feature selection from VFDT to keep decline accuracy constant.

REFERENCES

[1] J. Ross Quinlan. *C4.5 : programs for machine learning*, Morgan Kaufmann, San Mateo, Calif., 1993.

[2] P. Domingos. G. Hulten. Mining High-Speed Data Streams, *Proceedings of the ACM Sixth International Conference on Knowledge Discovery and Data Mining*, ACM Press, pp.71-80,2000.

[3] T. Minegishi. M. ISE. A. Niimi. O. Konishi. A proposal of abusing credit cards detecting systems using attribute selection method with multistage decision tree construction, *Information Processing Society of Japan*, The 71st National Convention of IPSJ, pp.603-604,2009.

[4] T. Minegishi. M. ISE. A. Niimi. O. Konishi. Comparison with two attribute selection methods using actual data, stepwise procedure in logistic regression analysis and selection by decision, *Japan Society for Fuzzy Theory and Intelligent Informatics*, The 25th Fuzzy System Symposium, 1A2-02 (6 pages in CD-ROM),2009.

[5] P. Domingos. G. Hulten. VFML - a toolkit for mining high-speed time-changing data streams, *http://www.cs.washington.edu/dm/vfml/*, 2003.

[6] Ian. H Witten. Eibe. Frank. *DATA MINING*, Morgan Kaufmann, pp.187-199, 2005.