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

Feature Selection in Large Scale Data Stream for Credit Card Fraud Detection

Masayuki Ise, Ayahiko Niimi, Osamu Konishi
Future University Hakodate
2, 116 Kameda-nakano-cho, Hakodate, Hokkaido, Japan
{g3108001, niimi, okonishi}@fun.ac.jp

Abstract — **There is increased interest in accurate model acquisition from large scale data streams. In this paper, because we have focused attention on time-oriented variation, we propose a method contracting time-series data for data stream. Additionally, our proposal method employs the combination of plural simple contraction method and original features. In this experiment, we treat a real data stream in credit card transactions because it is large scale and difficult to classify. This experiment yields that this proposal method improves classification performance according to training data. However, this proposal method needs more generality. Hence, we'll improve generality with employing the suitable combination of a contraction method and a feature for the feature in our proposal method.**

I. INTRODUCTION

The data stream (such as network communication log, transaction log and operation log) occurs randomly and has changed its feature day by day. The cumulative data stream is large scale. In addition, it has attracted the attention because achieving the best model is the most important.

For improvement of the prediction performance and the cost performance, part of the process for achieving the best model, new features are constructed or feature subset is selected. However, it's not easy that constructing feature newly without loss of generality and selecting the best feature subset when there are many features. For this reason, various researches are proposed in these days.

There are some studies such as method to select some similarities among features [1], a method to select dependence from among features to objective variable [2], converts feature of the data into the score and classify it by various techniques [3], to select features by changing a penalty dynamically [4]. In addition, there is a report of category, basic method and their problem about feature selection [5].

Using Forward-backward stepwise selection (stepwise) in this paper selects from dependency to objective variable also in [2]. However, using in the real-data, False Positive Rate (FPR) using in [2] is unable to function because quantitative balance between positive-class and negative-class is partial (FPR is always close enough to 100% when training data was classified at random). In a part of the process of stepwise, remove similar features by correlation coefficients between each feature such as [1]. Method in [4] for feature selection is useful when quantity of training data is not enough, however it was reported that is not useful when quantity of training data is large scale such as our research.

In a study of feature variation by feature subset selection bias [6], it's shown to experimental data that the feature variation is small when quantity of data is large scale.

In a study to classify fraud-use and normal-use for fraud detection from credit card transactions, best combination of training data sampling and classifies are extracted. Then it was best combination when quantities of fraud-use and normal-use were equivalent. Consequently, it warranted sampling the normal-use data in this paper. Treated data in this paper is difficult to classify more than the data in [7] because it is larger and more features than data in [7], and percentage of fraud-use data is 0.38% (percentage is 20% in [7]).

Because data stream keeps changing, we have focused attention on time-oriented variation. For this reason, we have employed some constructed features that were generated from some past data and were formulated by expert's knowledge. Those constructed features are called "behavior features" (refer to Chapter III). However, in a research of data stream mining, using behavior features with their some past data more as explaining variables was improved performance of classifier [8]. However when using past data as explain variables simply increase explaining variables greatly, it needs more computational cost for the analysis. For this reason, we propose using behavior features with contracting their past data more.

For improvement of performance of classifier in real data stream, the aim of this paper is to propose a best selection method that selects features from contracted features by managing plural time-oriented information contraction method.

In addition, we propose employing the suitable combination of a contraction method and a feature for the feature.

II. PROPOSAL METHOD

For feature construction, there are some approaches to construct a feature newly from different features and to extract a new feature from a part of a feature. Because we noted that the feature of data stream varies with time, we contracted a feature and its past features. In this way, we hypothesize that classifier will be improved performance by selecting effective features. We already have employed the feature that was formulated by the expert. That feature is

called “behavior feature” and generated by contracting past features (refer to Chapter III). However, this proposal method does not need the expert. We represent a behavior feature in the scholarly definition because it’s elaborate calculation (refer to Chapter III).

The procedure of this proposal method; 1) Constructing new feature from several features by time-oriented information contraction methods. 2) Selecting beneficial features for performance of classifier from among generated features by a procedure in 1). 3) Applying 1) and 2) to each features.

Let R be a set of possible features, $x_n (\in R)$ be a given feature. Let $X_n = \{x_{n,-1}, x_{n,-2}, \ldots, x_{n,-N}\}$ be a feature set of the past features of x_n. Let $Model$ be a set of feature for classifier. Let ϕ be an information reduction method, and $\phi(x_n, X_n)$ be a contracted feature that is generated by ϕ. Let $FS(\ )$ be a selection method. A FIGURE I is the algorithm of proposal method.

Initialize:

$R = \{x_1, x_2, \ldots, x_N\}$

$Model = \{\ \}$

Do each original feature: $x_n \in R$

/* X_n is past features of x_n */

$\Phi_n = \{x_n, \phi_1(x_n, X_n), \phi_2(x_n, X_n), \ldots, \phi_m(x_n, X_n)\}$

$Z_n = FS(\Phi_n)$

$Model = Model \cup Z_n$

Output: *Model*

FIGURE I: THE ALGORITHM OF PROPOSAL METHOD

Beneficial features are selected by $FS(\ )$ from x_n and $\{\phi_1(x_n, X_n), \phi_2(x_n, X_n), \ldots, \phi_m(x_n, X_n)\}$. In addition, those are added to $Model$. Then each feature is applied. In this paper, we use stepwise method as $FS(\ )$.

A FIGURE II is a flow of generating a classifier model with this proposal method. After training data was applied to sampling process, it constructs behavior features that are formulated by the expert. In addition, information contraction features are appended. Then it applies to feature selection process and generate classifier model.

A. *Credit Card Transaction Data*

Using credit card transaction data has occurred on real-time when a card-holder makes a payment by credit card. The scale of the quantity of data is large so that quantity of data to deal with in a finance company is beyond 1,000,000 transactions per day. Additionally, a couple hundreds of data occur in the peak for one second. Because the credit card is used all over the world, transactions occur 24-7. There are business information (such as Card ID, Transaction amount, Terminal ID) and customer information (such as Credit Limit) in these features.

Because credit card transaction data is large, complicated and difficult real time-series data that has 0.38% fraud-use data (and 53 original features), this data is well-suited for the validation of our proposal method.

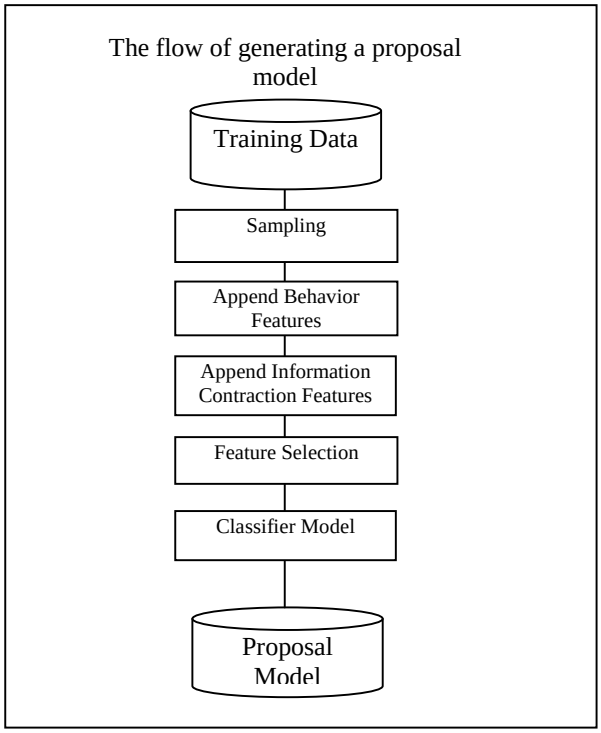

FIGURE II: THE FLOW OF A GENERATED PROPOSAL MODEL

III. BEHAVIOR FEATURES

The feature that was constructed separately from the original feature is called "processing feature". A transaction data (m-dimensional) is $R_0 = \{x_{1,0}, x_{2,0}, \ldots, x_{m,0}\}$.

The past data of R_0 before the *n*th transaction is $R_{-n} = \{x_{1,-n}, x_{2,-n}, \ldots, x_{m,-n}\}$. The constructed feature that was constructed from the feature of $\{R_{-1}, R_{-2}, \ldots, R_{-N}\}$ is called “behavior feature”. These features are used to analyze credit card transactions for fraud detection in financial company, and improve classification performance there. These include some unique methods that were formulated for quantifying expert’s knowledge or know-how.

The experts of credit card fraud detection have known that criminals behave strangely compared with each card holder’s

measure (such as buying jewelry goods abruptly, using a credit card when its holder is sleeping). According to these expert's knowledge, we formulated behavior features. However because expert's knowledge are various, we could not quantify easily. Together with experts we spend approximately a year for formulating behavior features. In actual classification work for fraud detection, these unique methods have employed and have improved the performance of classifier.

In this paper, we introduce *Round-use Coefficient*. Based on expert's knowledge or know-how, this coefficient quantifies that how many times criminals used credit card at short period of time. This method replaces expert's knowledge (such as if a credit card was used n times in t hours.) to a window function. Past features are replaced a behavior feature by using that windows function. This feature expresses the bigger the value is more dubious. We illustrate an example concretely.

TABLE I
THE EXAMPLES OF THE EXPERT'S KNOWLEGDE FOR ROUND-USE COEFFICIENT

THE NOMBER OF KNOWLEDGE	WITHIN *T* HOUR(S)	OVER *C* TRANSACTION(S)
1	$T_1 = 0.25$	$C_1 = 1$
2	$T_2 = 1.00$	$C_2 = 3$
3	$T_3 = 1.50$	$C_3 = 4$

A. An Example of "Round-use Coefficient"

Experts of credit card transaction represent the feature of round-use by the number of times per time. For example, there is expert's knowledge such as TABLE I. Let $w_2(t_{-n})$ be a window function of second knowledge, denote following a formula.

$$w_2(t_{-n}) = \begin{cases} f(t_{-n}), & T_1 \le t_{-n} < T_2 \\ 0, & otherwise \end{cases} \quad (1)$$

Let x_0 be a current credit card transaction, x_{-n} be n times earlier transactions of x_0. Let t_{-n} be time interval between a time of x_0 occurred and a time of x_{-n} occurred. Let $f(\ )$ be a monotone decreasing function (such as linear function or sigmoid function) that pass through the points of (T_1, C_1^{-1}) and (T_2, C_2^{-1}). Let $w_k(t_{-n})$ be a window function of k th knowledge, denote following a formula when $f(\ )$ means linear function (Slope and intercept are (a/d) and (b/d)).

$$w_k(t_{-n}) = \begin{cases} -\frac{a}{d}t_{-n} + \frac{b}{d}, & T_{k-1} \le t_{-n} < T_k \\ 0, & otherwise \end{cases} \quad (2)$$

$$\left.\begin{aligned} a &= C_k - C_{k-1} \\ b &= C_k T_k - C_{k-1} T_{k-1} \\ d &= C_k C_{k-1}(T_k - T_{k-1}) \end{aligned}\right\} \quad (3)$$

A value of Round-use Coefficient is obtained following a formula from the past transactions. Let N be the number of past transactions. We define K as the number of knowledge (in example, $K = 3$)

$$W = \sum_{n=1}^{N} \sum_{k=1}^{K} w_k(t_{-n}) \quad (4)$$

IV. CONTRACTION METHODS

We explain the method that we employed for an information contraction in this chapter. It is noted that using original features, behavior features and some past data achieves more than using these features without some past data [8]. However, appending simply past data needs more computational cost because it mounts up many input features. By contracting some past behavior features to a new feature, it solves this problem. Because current and past feature sets are able to regard the time-series, those new features retain feature of original data by contracting the time-series variation, the statistical feature and the clustering of dataset.

There are three main possible ways in the reduction about time-series data.

- Describing the bias of time-series data by approximation formula.
- Replacing time-series data with statistical representative value.
- Focusing on the similarity and the bias of distribution in time-series data.

In this paper, we implement a number of methods with each of these ways. Additionally, we evaluated these methods to improvement of classifier.

TABLE II
THE INFORMATION OF THE TREATED DATA

INFO.	VALUE	NOTE
Quantity of data	13,583,331	Internalize fraud-use 51,984
Percentage of fraud-use	0.38%	51,984÷13,583,331
Features	53	Internalize 29 behavior features

Those are categorized in *Straight-Line Approximation*, *Representation Value* and *Similarity*. Because we used large-scale data stream, we employed some methods with low cost. We employed the simple method that did not need an expert

knowledge unlike the behavior feature in Chapter III.

We propose employing the suitable combination of a contraction method and a feature for the feature.

A. *Straight-Line Approximation*

Let t be the time when x_n has occurred, denoting $X\left(=\left\{x_n, X_n\right\}\right)$ and $T\left(=\left\{t, t_{-1}, \ldots, t_{-N}\right\}\right)$. The constructed features calculate a slope and an intercept by $X = a \times T + b$. These features describe the time series variation by a straight line.

B. *Representative Value*

Each of these methods, we choose the representative value from a feature and the past features. In this paper, we employed arithmetic mean, the maximum, the minimum and the range.

C. *Similarity*

These methods calculate a distance of a feature and the past features. We employed the Euclidean norm, the similarity and the extended similarity.

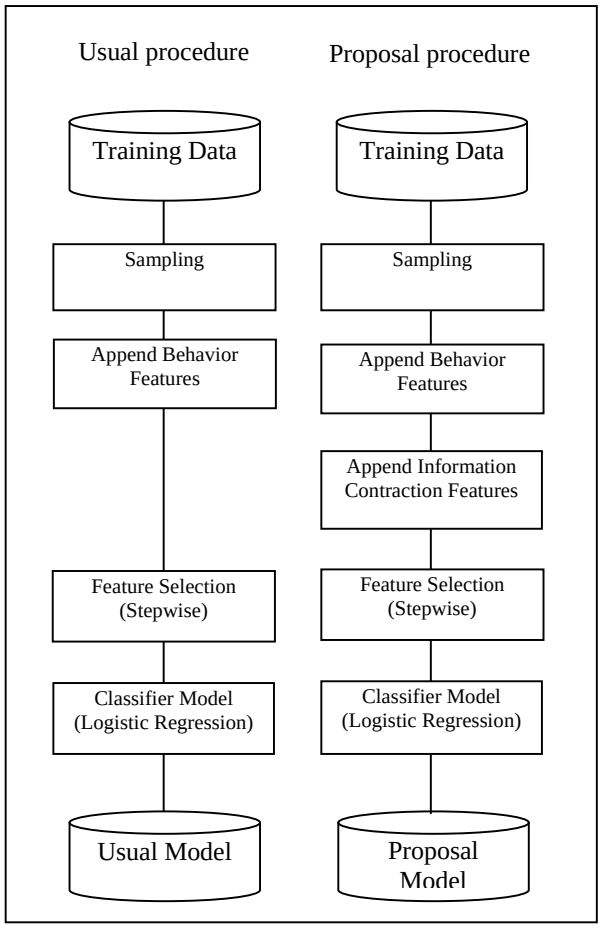

FIGURE III: USUAL PROCEDURE AND PROPOSAL PROCEDURE

V. EXPERIMENT

In this experiment, we treated real credit card transaction data that have 53 original features (internalize 29 behavior features) and have 13,583,331 observations (internalize 51,984 fraud-use observations). We compared the classification performance of the generated model by a proposal procedure and the generated model by a procedure that we employed so far. In this comparison, we confirmed stability by 10 times operation.

FIGURE III is the flow of proposal procedure and the flow of usual procedure. Append Behavior Features adds the behavior feature (refer to Chapter III) that the expert formulated with us. We employed Stepwise to the feature selection. We generated a Classifier Model in logistic regression and we made scoring by that model and classified credit card transaction data.

Append Information Contraction Features adds the feature that contracts the past data of each feature (refer to Chapter IV).

In actual classification work for fraud detection, because credit card transaction data is large scale, operators can process only 0.2% of all transactions. Consequently, we evaluated the performance of classifier with top 0.2% of the cap-curve.

A. *Experimental Results*

We confirmed relations between the behavior feature and the rate of fraud-data. Thence, we exhibit a most interested result in FIGURE IV.

FIGURE IV shows logit of fraud-rate according to Round-use Coefficient in this experimental data. Fraud-rate means a rate of fraud-data.

Therefore, in this experimental data, Round-use Coefficient describes the relation to fraud-rate.

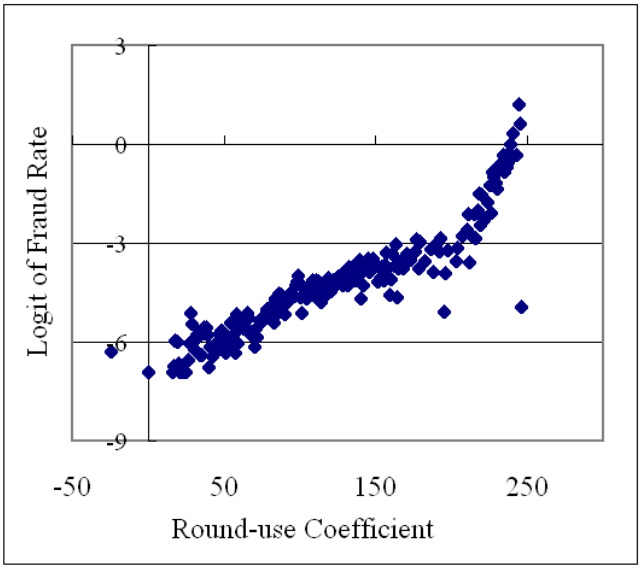

FIGURE IV: LOGIT OF FRAUD-RATE BY ROUND-USE COEFFICIENT

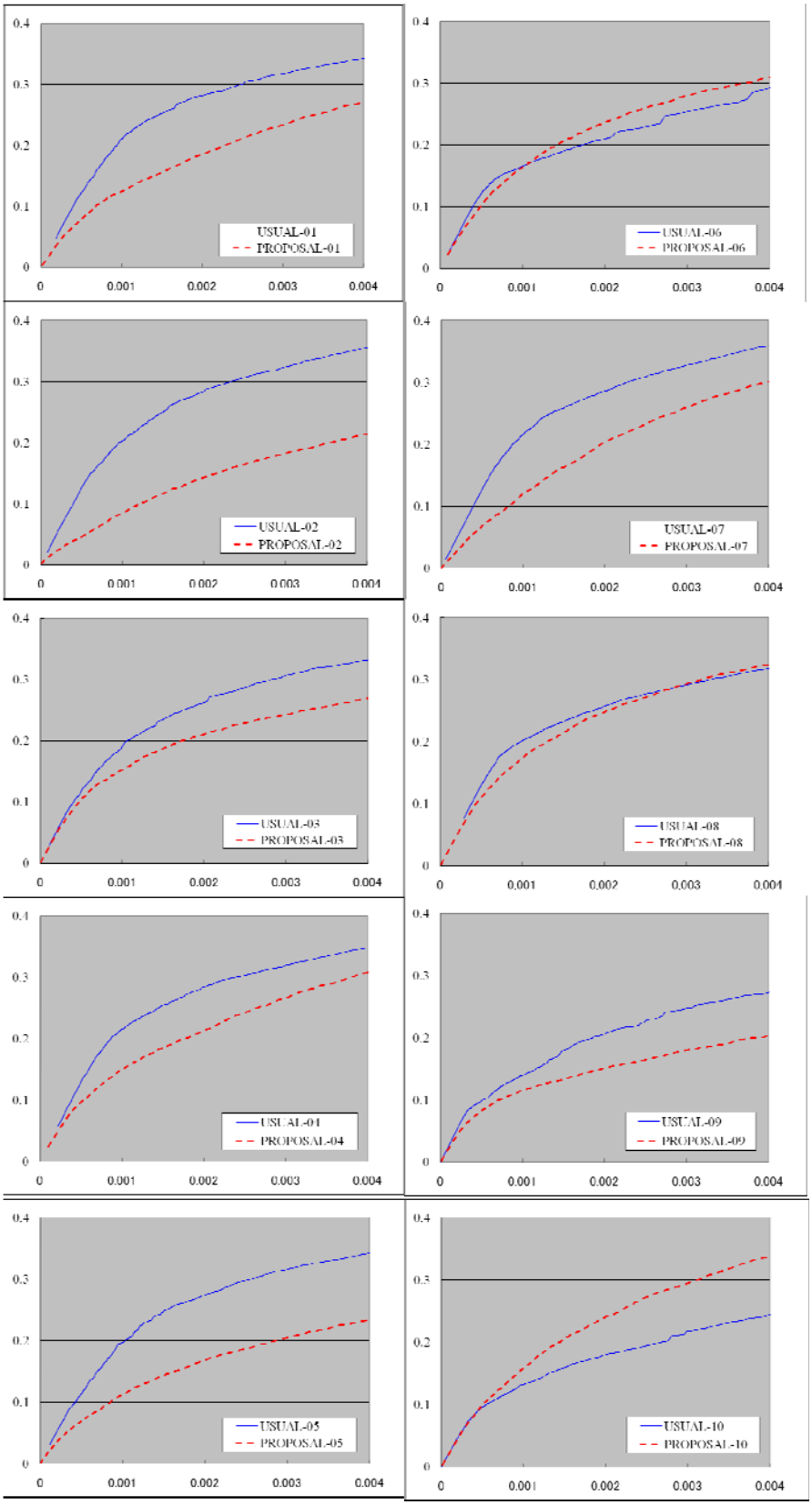

FIGURE V: CAP CURVES OF USUAL MODEL AND PROPOSAL MODEL

FIGURE V is experimental result. Horizontal axis is denoted a rate of transaction according to the highest-ranking score. Vertical axis is denoted a rate of detected fraud-transaction of all fraud-transaction.

It is because operators cannot process only 0.2% of all transaction, the point of focus in FIGURE V is a point of 0.002 on cap curves. In nine cases out of ten cases, usual models are the same or more than proposal models. However, in other 1 case, classifier model was improved. Due to a number of selected features ware relatively small, the one case was different from the rest. In this experiment, this proposal procedure does not have stability.

Common features in all usual models were 17 features. Each usual model selects nearly 40 features. In contrast, there was only one common feature, in all proposal models. Each proposal model selects nearly 70 features. For this reason we consider that the performance of our proposal models had not stability. Moreover, those features which have many missing values were not selected in this proposal selection method. In addition, because this proposal selection method generated a number of features, it might have a adverse impact (such as curse of dimensionality) on proposal models.

Accordingly, we have considered that this problem will be solved by defining the combination of contraction method and feature according to the feature. Consequently, our proposal method will be improved because it reform from the reducing to generate features and the using some appropriate methods for each feature.

VI. SUMMARY

For the data stream, we propose a selection method for improving performance of the classifier that a selection method employs time-oriented information contraction method to feature construction.

Using real credit card transaction data for a month, we compared the classification performance with top 0.2% of the cap-curve by 10 times operation. This experiment yields that this proposal method improves classification performance according to training data. However, this proposal method needs more generality.

In this proposal selection method, possible features for a model are nearly 500 features because this method searches the best combination of the feature and the information contraction method from all possible combinations. Henceforth, we'll reduce the combination selection for possible model, and improve the generality by the beneficial change of proposal selection method.

REFERENCES

[1] Z. Zhao and H. Liu: Spectral Feature Selection for Supervised and Unsupervised Learning. *Proc. 24th Int. Conf. on ML*, 1151-1158. (2007)

[2] L. Song, A. Smola, A. Gretton, K.M. Borgwardt and J. Bedo: Supervised Feature Selection via Dependence Estimation. *Proc. 24th Int. Conf. on ML*, 823-830. (2007)

[3] J. Zhang and Y. Yang: Probabilistic Score Estimation with Piecewise Logistic Regression. *Proc. 21st Int. Conf. on ML*, 115-122. (2004)

[4] J. Zhou, D. Foster, P. Stine and L. Ungar,: Streaming Feature Selection using Alpha-investing. *KDD'05*, 384-393. (2005)

[5] I. Guyon and A. Elisseeff: An Introduction to Variable and Feature Selection. *JMLR.*, 3, 1157-1182. (2003)

[6] S. Singhi and H.Liu: Feature Subset Selection Bias for Classification Learning. *Proc. 23rd Int. Conf. on ML*, 849-856. (2006)

[7] Stolfo, S.J., Fan, D.W., Lee, W., Prodromidis, A.L., and Chan, P.K.: Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results. *Working notes of AAAI Workshop on AI Approaches to Fraud and Risk Management*, 83-90. (1997)

[8] M. Tsuzuki and O. Konishi: Adaptive Kernel Methods in Large-scale Data Stream with Historical Information. *Trans. IPS Japan, Vol.1, No.1*, 49-59. (2008)