PAPER

Risk-sensitive Learning via Minimization of Empirical Conditional Value-at-risk

Hisashi KASHIMA[†], *Member*

SUMMARY We extend the framework of cost-sensitive classification to mitigate risks of huge costs occurring with low probabilities, and propose an algorithm that achieves this goal. Instead of minimizing the expected cost commonly used in cost-sensitive learning, our algorithm minimizes conditional value-at-risk, also known as expected shortfall, which is considered a good risk metric in the area of financial engineering. The proposed algorithm is a general meta-learning algorithm that can exploit existing example-dependent cost-sensitive learning algorithms, and is capable of dealing with not only alternative actions in ordinary classification tasks, but also allocative actions in resource-allocation type tasks. Experiments on tasks with example-dependent costs show promising results.

key words: *risk-sensitive learning, cost-sensitive learning, meta learning, conditional value-at-risk, expected shortfall*

1. Introduction

Classification learning is one of the fundamental tasks in data mining. It is widely seen in many important tasks in the real world such as diagnostics in health care, credit administration in finance, campaign management in direct marketing, and so on. Its task is to predict the actions (or classes) of the target objects whose appropriate action (or classes) are unknown given pairs of an object and its appropriate action (or class) as training examples. In other words, it aims to minimize the probability of misclassification.

However, there are many cases where it is not enough only to minimize the number of mistakes. For example, the cost of misdiagnosis of classifying healthy people as sick and that of classifying sick people as healthy are apparently not equal, since the latter leads to serious results. Moreover, the degree of seriousness differs among patients.

Similarly, when we make management decisions on what project we should invest in, the execution cost, profit from success, and loss from failure depend on the characteristics of the project.

Cost-sensitive learning [1]–[8] is a suitable framework for such cases where costs are different among classes or objects, and the amounts of them are unknown at the stage of prediction. Wider range of problems can be treated in the framework since it aims to minimize not the probability of misclassification but the expected cost of misclassification. The ordinary classification problem is understood as a special case that assumes that all costs of misclassification are 0 or 1.

However, there can be situations where cost-sensitive learning is still not enough. Since minimizing the expected cost does not aggressively suppress the occurrence of huge costs, it can not avoid such a risk of disasters. Therefore, if there is not a little chance of huge costs, and also if users are interested in mitigating such risk, minimization of the expected cost does not reflect the objective. Actually, risk aversion is one of the central topics in financial engineering. For example in portfolio theory, the analyst is expected to find a portfolio that maximizes profit while suppressing the risks of huge costs occurring with low probabilities [9].

In this paper, we define the term *risk* as *chances of huge costs occuring even with small probability*, and propose an approach of *risk-sensitive classification learning* that considers cost distributions not to decrease the expected cost but to mitigate the risks of huge costs. Concretely, instead of the expected cost, we employ a risk metric called conditional value-at-risk (CVaR) [10], also known as expected shortfall, which is attracting considerable attentions in financial engineering. We propose a risk-sensitive learning algorithm that minimizes the CVaR as the objective function. Our algorithm is a meta-learning algorithm, which is quite a general procedure that can convert existing cost-sensitive learners to risk-sensitive learners.

This paper is organized as follows. In Section 2, we review the definition and the existing approaches of cost-sensitive learning, and then point out a drawback of these approaches. In Section 3, we introduce our risk-sensitive learning approach using CVaR as the objective function, and propose a meta learning algorithm, MetaRisk. We also introduce reduction from risk-sensitive learners to cost-sensitive learners, not only with alternative actions, but also with allocative actions that are not considered in ordinary cost-sensitive learning problems. In Section 4, we show some experimental results on two datasets, a synthetic dataset and a more realistic dataset for credit administration. In Section 5, we review related works, and discuss their relations to our approach. Finally, we conclude this paper with discussion and future work.

Manuscript received January 1, 2003.
Manuscript revised January 1, 2003.
Final manuscript received January 1, 2003.

[†]The author is with the IBM Research, Tokyo Research Laboratory.

2. Cost-Sensitive Learning

In this section, we review the definition and the existing approaches of cost-sensitive learning, especially with example-dependent costs. And then, we point out a drawback of these approaches from the standpoint of risk aversion.

2.1 Decision Model

Let X be a set of all *target objects*, for example $X = \mathbb{R}^M$, and Y be a finite set of *actions* taken against the target objects. For example in the context of direct marketing, $\mathbf{x} \in X$ is a customer profile, and Y is a set of possible marketing actions such as direct mail, email, telemarketing, and so on.

Function h is called *hypothesis*, and defined as $h(\mathbf{x}, y; \theta) : X \times Y \to \mathbb{R}$, where θ is its model parameters. An action $\hat{y} \in Y$ taken against $\mathbf{x} \in X$ is determined by

$$\hat{y} = \operatorname*{argmax}_{y \in Y} h(\mathbf{x}, y; \theta). \tag{1}$$

Usually, only one action is assumed to be taken at a time, hence we call these types of actions *alternative actions*.

We might assume the following stochastic constraint in $h(\mathbf{x}, y; \theta)$,

$$\sum_{y \in Y} h(\mathbf{x}, y; \theta) = 1, \text{ s.t. } h(\mathbf{x}, y; \theta) \geq 0, \tag{2}$$

for $\forall \mathbf{x} \in X, \forall y \in Y$. Instead of (1), we can make stochastic selection of one alternative actions with probability distribution (2).

If it is allowed to take multiple actions at a time, and to allocate resources to each of $|Y|$ actions in proportion to $h(\mathbf{x}, y; \theta)$ with (2), those kind of actions are called *allocative*. Allocative actions are popular in the context of portfolio selection [9] where funds are allocatively invested to financial products.

In this paper, we deal with those two cases, in one of which an action is alternatively chosen with (1), and in the other of which stochastic selection or resource allocation is allowed with (2).

2.2 Cost

Cost is a random variable $c(\mathbf{x}, y) \in \mathbb{R}$ defined over $X \times Y$, which indicates how bad an action $y \in Y$ taken against $\mathbf{x} \in X$ is.[†].

For instance in medical diagnosis, $c(\mathbf{x}, y)$ is the badness of the medical treatment y taken for a patient with the results of medical tests $\mathbf{x}$. $c(\mathbf{x}, y)$ becomes small if the treatment is appropriate, and becomes large if not. If the treatment is significantly inappropriate, and his or her health is lost, $c(\mathbf{x}, y)$ becomes huge.

In this paper, we deal with the most general problem setting in cost-sensitive learning, where the true cost distribution is unknown, and depends on examples [5], [6], [8]. Note that although those literatures assume that the cost also depends on classes, we adopt the notation without the dependency since it is convenient to think that the cost incorporates the dependency on classes implicitly.

Also, following the context of cost-sensitive learning, we evaluate actions in terms of cost instead of reward or profit, but the following discussion still holds for reward or profit by simply changing those signs.

Let $c(\mathbf{x}, h(\theta))$ be the cost of the action for $\mathbf{x}$ by using hypothesis $h(\mathbf{x}, y; \theta)$. In the case of alternative actions (1), $c(\mathbf{x}, h(\theta))$ becomes

$$c(\mathbf{x}, h(\theta)) = c(\mathbf{x}, \operatorname*{argmax}_{y \in Y} h(\mathbf{x}, y; \theta)). \tag{3}$$

In the case of allocative actions, it is not trivial to represent $c(\mathbf{x}, h(\theta))$. We consider the simplest case where $c(\mathbf{x}, h(\theta))$ is represented as

$$c(\mathbf{x}, h(\theta)) = \sum_{y \in Y} h(\mathbf{x}, y; \theta) c(\mathbf{x}, y), \tag{4}$$

where the cost of each action linearly depends the amounts of investment to the action. This form corresponds to the return of a portfolio used in portfolio theory [9].

Note that if we make stochastic selection of an alternative action by (2), we can also use (4), but this is not the realized cost, but the expected cost for $\mathbf{x}$.

2.3 Cost-Sensitive Learning

Cost-sensitive learning [1]–[8] is a framework for supervised classification learning with cost $c(\mathbf{x}, y)$. In cost-sensitive learning, the expected cost is conventionally used as the objective function for training to find the best θ. The expected cost with respect to data distribution D over $X \times \mathbb{R}^Y$ is defined as

$$C^D(\theta) = E_D \left[\, c(\mathbf{x}, h(\theta)) \, \right]. \tag{5}$$

Unfortunately, since we do not know D, we exploit training examples E instead. N training examples in E are assumed to be independently sampled from D. Let the i-th training example in E be $\mathbf{e}^{(i)} = (\mathbf{x}^{(i)}, \{c^{(i)}(\mathbf{x}^{(i)}, y)\}_{y \in Y})$, where $\mathbf{x}^{(i)} \in X$ is the i-th target object and $c^{(i)}(\mathbf{x}^{(i)}, y)$ is the cost of action $y \in Y$ for $x^{(i)}$. Note that the cost of every action is given for each training example.

The empirical expected cost for the training examples is defined as

$$C^E(\theta) = \frac{1}{N} \sum_{i=1}^{N} c(\mathbf{x}^{(i)}, h(\theta)). \tag{6}$$

[†] We do not care whether or not there are influential factors $c(\mathbf{x}, y)$ for other than x and y.

Since $C^E(\theta)$ is a good approximation of $C^D(\theta)$ for sufficiently large N, parameter θ is determined so that $C^E(\theta)$ is minimized [5], [6], [8].

2.4 Drawback of Mean-Cost Minimization Approach

Let us imagine a situation where occurrences of huge costs are fatal. For example, if we have to make important management decisions, several consecutive mistaken judgements might directly lead to bankruptcy. Also, if the costs follow heavy-tailed distributions, the expected cost is highly affected by one big cost. In those cases where there are chances of unacceptably huge costs occurring even with small probability, one would like to avoid those risks as far as possible. In this paper, we define the term *risk* as *chances of huge costs occuring even with small probability.*

Let us consider another example. Assume that two hypotheses $h(\theta_1)$ and $h(\theta_2)$, and both of them have identical expected costs. Consider two hypotheses $h(\theta_1)$ and $h(\theta_2)$, both of them having identical expected costs. $h(\theta_1)$ has a cost distribution with high peak around its expected cost, and $h(\theta_2)$ has one with a gentle slope and a heavy tail in its high cost area. In this situation, risk aversive investors would apparently prefer $h(\theta_1)$ to $h(\theta_2)$.

The above discussion implies that minimization of the expectation of $c(\mathbf{x}, h(\theta))$ is not enough, and suggests the need to consider the distribution of $c(\mathbf{x}, h(\theta))$ and aggressively avoid the risk of huge costs.

3. Risk-Sensitive Learning

Motivated by the discussion in the previous section, we define *risk-sensitive learning* as approaches using risk metrics as the alternative objective functions that aggressively avoid huge costs instead of the expected cost, and propose to use the conditional value-at-risk (CVaR) as our objective function. And then we propose a meta-learning algorithm that reduces cost-sensitive learners to risk-sensitive learners to minimize CVaR.

3.1 Risk-Sensitive Learning via Minimizing Conditional Value-at-risk

3.1.1 Value-at-Risk

In the area of financial engineering, various risk metrics such as standard deviation, beta, and Sharpe ratio [9] have been studied for decision making with low risk of huge costs. Probably, one of the most popular risk metrics is value-at-risk (VaR) [12]. Value-at-risk is defined to be the β-quantile of the cost distribution for a given constant $0 \le \beta \le 1$. In other words, it is the minimum of the top $100(1-\beta)\%$ costs. In our problem setting, the value-at-risk $\alpha_\beta^D(\theta)$ with respect to hypothesis h and data distribution D is defined as (See Figure 1.)

$$\alpha_\beta^D(\theta) = \min \left\{ \alpha \in \mathbb{R} \;\middle|\; E_D \left[I \left(c(\mathbf{x}, h(\theta)) \ge \alpha \right) \right] \le 1-\beta \right\},$$

where $I(\cdot)$ is a function that returns 1 when its argument is true, and returns 0 otherwise. Note that the value-at-risk depends on model parameters θ.

Although value-at-risk is a widely-accepted risk metric, some drawbacks have been pointed out [12]. One problem is that once the cost surpasses the value-at-risk, it is not cared at all how huge the cost becomes. On the other hand, we are rather interested in suppressing the amount of huge costs itself. Also, value-at-risk has been shown to be non-convex in most cases theoretically and empirically, which is extremely inconvenient. If the cost distribution follows a Gaussian distribution, the value-at-risk becomes a linear combination of the mean and the standard deviation of the cost, and the above problems are resolved. However, the assumption usually does not hold.

3.1.2 Conditional Value-at-risk (CVaR)

Conditional value-at-risk (CVaR) [10], also known as expected shortfall, is attracting attentions as a relatively new risk metric in the field of financial engineering. It is defined as the expected costs above the value-at-risk, in other words, the expectation of the top $100(1-\beta)\%$ costs (See Figure 1.), hence it can consider the amount of huge costs. Moreover, CVaR has desirable characteristics such as convexity [13]. This is exactly the risk metric that we want to employ as the objective function of risk-sensitive learning.

In our problem setting, the CVaR $\phi_\beta^D(\theta)$ with respect to hypothesis h and data distribution D is defined as

$$\phi_\beta^D(\theta) = \frac{1}{1-\beta} E_D \left[I \left(c(\mathbf{x}, h(\theta)) \ge \alpha_\beta^D(\theta) \right) \cdot c(\mathbf{x}, h(\theta)) \right], \tag{7}$$

where $\alpha_\beta^D(\theta)$ is the value-at-risk defined above, and note that the definition of th CVaR uses value-at-risk.

Since the CVaR can be seen as the conditional expected cost surpassing $\alpha_\beta^D(\theta)$, (7) is decomposed into two terms as

$$\phi_\beta^D(\theta) = \alpha_\beta^D(\theta) + \frac{1}{1-\beta} E_D \left[c(\mathbf{x}, h(\theta)) - \alpha_\beta^D(\theta) \right]^+, \tag{8}$$

where $[x]^+$ is a function that returns x when $x \ge 0$, and returns 0 otherwise.

3.2 Model Estimation

3.2.1 MetaRisk: A Risk-Sensitive Learner to Minimize CVaR

Let us derive an algorithm to optimize parameter θ.

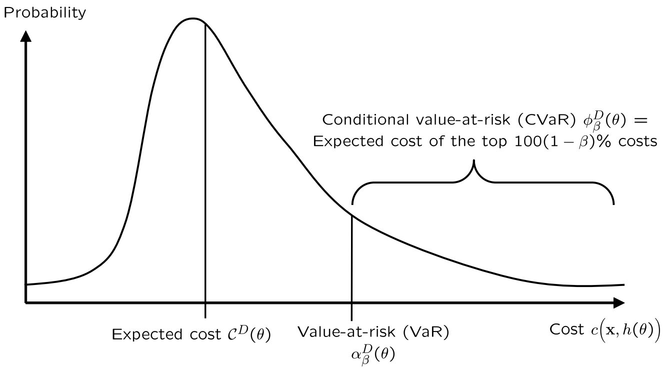

Fig. 1 Expected cost, value at risk (VaR), and conditional value-at-risk (CVaR).

Although (8) is the objective function that we want to minimize, we employ the following empirical CVaR defined on training examples E instead of D which is unknown.

$$\phi^E_\beta(\theta) =$$

$$\alpha^E_\beta(\theta) + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} \Big[\ c(\mathbf{x}^{(i)}, h(\theta)) - \alpha^E_\beta(\theta)\ \Big]^+, \quad (9)$$

where $\alpha^E_\beta(\theta)$ is the value-at-risk for the training examples E,

$$\alpha^E_\beta(\theta) = \quad (10)$$

$$\min \left\{\ \alpha \in \mathbb{R}\ \middle|\ \frac{1}{N} \sum_{i=1}^{N} I\ \big(\ c(\mathbf{x}^{(i)}, h(\theta)) \geq \alpha\ \big) \leq 1 - \beta\ \right\}.$$

Now, if we suppose that $\alpha^E_\beta(\theta)$ is a known constant $\tilde{\alpha}$ in (9), we only have to minimize the second term (9),

$$\tilde{C}^E_{\tilde{\alpha}}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \Big[\ c(\mathbf{x}^{(i)}, h(\theta)) - \tilde{\alpha}\ \Big]^+. \quad (11)$$

Note that (11) is convex if $c(\mathbf{x}^{(i)}, h(\theta))$ is convex with respect to θ. For the time being, we assume existence of algorithms to find θ that minimizes (11).

Next, we fix θ, and find the VaR (10) for θ. Since (10) is defined for the training examples E, it is rewritten as

$$\alpha^E_\beta(\theta) = \min_{k=1,\ldots,N} \Bigg\{\ c(\mathbf{x}^{(k)}, h(\theta))\ \Bigg|$$

$$\frac{1}{N} \sum_{i=1}^{N} I\Big(c(\mathbf{x}^{(i)}, h(\theta)) \geq c(\mathbf{x}^{(k)}, h(\theta))\Big) \leq 1 - \beta\ \Bigg\}$$

which is equivalent to $c(\mathbf{x}^{(k)}, h(\theta))$ where k is the index of the training datum with the $\lfloor (1-\beta)N \rfloor$-th largest cost by θ. $\alpha^E_\beta(\theta)$ is naively computed by sorting the costs by θ in $O(N \log N)$, or it can be reduced to $O(N)$ by using efficient algorithms for finding order statistics [14].

```
Algorithm: MetaRisk($E, \beta$)
[Step:1] Set $\tilde{\alpha} := 0$.
[Step:2] For the current $\tilde{\alpha}$, find $\theta' = \operatorname{argmin}_\theta \tilde{C}^E_{\tilde{\alpha}}(\theta)$,
         and set $\theta := \theta'$.
[Step:3] For the current $\theta$. find the empirical
         VaR $\alpha^E_\beta(\theta)$, and set $\tilde{\alpha} := \alpha^E_\beta(\theta)$.
[Step:4] Continue [Step:2] and [Step:3] until the
         convergence of $F^E_\beta(\theta, \tilde{\alpha})$.
```

Fig. 2 MetaRisk: Risk-sensitive meta-learning algorithm.

Based on the above discussion, we propose a risk-sensitive meta-learning algorithm named MetaRisk (Figure 2)[†], which minimizes the empirical CVaR by exploiting existing cost-sensitive learners, and by finding the model parameter and the corresponding value-at-risk alternately.

3.2.2 Optimality and Convergence of MetaRisk

The optimality and convergence of the algorithm (Figure 2) are directly guaranteed by the following theorem by [13] that shows the convexity of the upper bound of CVaR.

Theorem 1 (Rockafellar and Uryasev, 2000): Let

$$F^E_\beta(\theta, \alpha) = \alpha + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} \Big[\ c(\mathbf{x}^{(i)}, h(\theta)) - \alpha\ \Big]^+, \quad (12)$$

then

$$\min_\theta \phi^E_\beta(\theta) = \min_{\theta,\alpha} F^E_\beta(\theta, \alpha), \quad (13)$$

where $\phi^E_\beta(\theta)$ is the empirical CVaR defined in (9).

$F^E_\beta(\theta, \alpha)$ is convex with respect to α. If (6) is convex with respect to θ, $F^E_\beta(\theta, \alpha)$ is also jointly convex with respect to θ and α. Also,

$$\alpha^E_\beta(\theta) = \min \left\{\ \alpha \in \operatorname*{argmin}_\alpha F^E_\beta(\theta, \alpha)\ \right\} \quad (14)$$

holds. □

(13) indicates that minimization of (12) is equivalent to minimization of CVaR, and the joint convexity of (12) ensures the gradient-based optimization with respect to θ and α. Moreover, from (14), $\alpha^E_\beta(\theta)$ is the minimizer of $F^E_\beta(\theta, \alpha)$ at θ, hence MetaRisk exactly performs coordinate-wise descent of $F^E_\beta(\theta, \alpha)$.

3.3 Reduction from Risk Sensitive Learners to Cost-Sensitive/Insensitive Learners

3.3.1 Recycling Existing Cost-Sensitive Learners

In the previous subsection, we assumed to have learning algorithms to find θ that minimizes (11). However,

[†] MetaRisk is named after the cost-sensitive meta-learning algorithm MetaCost [3].

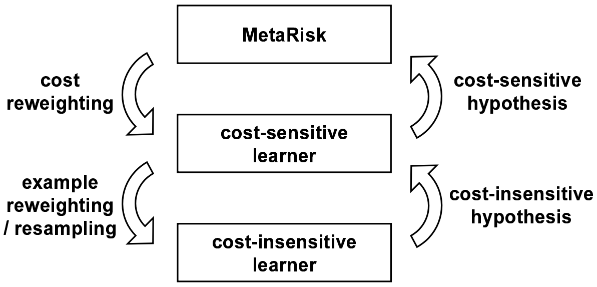
MetaRisk
cost reweighting
cost-sensitive hypothesis
cost-sensitive learner
example reweighting / resampling
cost-insensitive hypothesis
cost-insensitive learner

Fig. 3 Reduction from risk-sensitive learners to cost-insensitive learners.

it is not desired to design from scratch the risk-sensitive versions of existing learners such as perceptrons, decision trees, or SVMs. In this subsection, we demonstrate approaches that minimize (11) by iteratively calling existing example-dependent cost-sensitive learners with reweighted costs based on the current hypothesis.

Several example-dependent cost-sensitive learners [4], [7], [8] realize cost-sensitive learning by weighting or resampling training examples according to their costs, and feeding them to cost-insensitive learners. Merging this mechanism with our reduction enables converting existing cost-insensitive learners into risk-sensitive learners (See Figure 3).

3.3.2 Hypothesis with Alternative Actions

Reduction is relatively easy in the case of alternative actions (1). Paying attentions to its similarity to (6), we notice that this is the expectation of only costs exceeding $\alpha_\beta^E(\theta)$. Also, since actions are exclusive to each other, realized costs are limited to the form of $[c^{(i)}(\mathbf{x}^{(i)}, y) - \tilde{\alpha}]^+ + \tilde{\alpha}$. Therefore, substituting

$$\tilde{c}^{(i)}(\mathbf{x}^{(i)}, y) \;=\; [c^{(i)}(\mathbf{x}^{(i)}, y) - \tilde{\alpha}]^+ \tag{15}$$

for the original costs, (6) becomes

$$\tilde{C}_{\tilde{\alpha}}^E(\theta) = \frac{1}{N}\sum_{i=1}^{N} \tilde{c}^{(i)}(\mathbf{x}^{(i)}, y), \tag{16}$$

and this has the same form as the expected cost (6).

The reduction is realized by feeding example-dependent cost-sensitive learners [6]–[8] with modified training examples $\tilde{E}$, where the i-th example of $\tilde{E}$ is defined as

$$\tilde{e}^{(i)} \;=\; \left(\mathbf{x}^{(i)}, \{\tilde{c}^{(i)}(\mathbf{x}^{(i)}, y)\}_{y\in Y}\right).$$

3.3.3 Hypothesis with Allocative Actions

Next, let us consider the case where stochastic or allocative decision making by the constrained hypothesis (2) is allowed. (11) is rewritten as

$$\tilde{C}_{\tilde{\alpha}}^E(\theta) = \sum_{i=1}^{N}\Big[\sum_{y} h(\mathbf{x}^{(i)}, y; \theta) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha}\Big]^+ . \tag{17}$$

Unlike the case of alternative actions, $c^{(i)}(\mathbf{x}^{(i)}, h(\theta))$ depends on a convex combination of $c^{(i)}(\mathbf{x}^{(i)}, y)$, hence simple reweighting like (15) does not work.

A natural choice of the classifiers used as $h(\mathbf{x}, y; \theta)$ is the exponential family satisfying (2) such as multiclass logistic regression. However, in logistic regression, $c(\mathbf{x}^{(i)}, h(\theta))$ is not convex with respect to its parameters, and even worse, it is a multi-modal function. Therefore, we employ a family of classifiers with which $c(\mathbf{x}^{(i)}, h(\theta))$ is linear with respect to θ. (17) is convex with respect to its parameters. In this paper, we use gradient boosting [15], [16] as our optimization approach.

In gradient boosting, $h(\mathbf{x}, y; \theta)$ is represented as a linear combination of T deterministic hypotheses $f_1, \ldots, f_T$,

$$h(\mathbf{x}, y; \theta) = h_T(\mathbf{x}, y; \theta_T) = \sum_{t=1}^{T} w_t f_t(\mathbf{x}, y),$$

where $\theta_t = (w_1, \ldots, w_t)$ are the parameters. Since $h(\mathbf{x}, y; \theta)$ has to satisfy the stochastic constraints (2), we need

$$\sum_{t=1}^{T} w_t = 1, \text{ s.t. } w_t \geq 0.$$

At each boosting round t, suppose that we already have h_{t-1}, a new weak hypothesis f_t is sequentially added to h_{t-1} to construct h_t. h_t is recursively represented as

$$\begin{aligned} h_t(\mathbf{x}, y; \theta_t) &= (1-\gamma_t) h_{t-1}(\mathbf{x}, y; \theta_{t-1}) + \gamma_t f_t(\mathbf{x}, y) \\ &= h_{t-1}(\mathbf{x}, y; \theta_{t-1}) + \gamma_t (f_t(\mathbf{x}, y) - h_{t-1}(\mathbf{x}, y; \theta_{t-1})), \end{aligned}$$

where $0 < \gamma_t \leq 1$ is a updating parameter at round t, and finally, the parameters θ_t are determined as

$$w_t = \gamma_t \prod_{\tau=t+1}^{T} (1-\gamma_\tau).$$

Once f_t is determined, (17) is convex and piecewise linear with respect to γ_t. Therefore, γ_t is easily found by linear search or linear programming.

In order to find the weak hypothesis f_t at the boosting round t, assume that γ_t is sufficiently small, then the Taylor series expansion of (17) around h_{t-1} gives

$$\begin{aligned} \tilde{C}_{\tilde{\alpha}}^E &= \sum_{i=1}^{N}\Big[\sum_{y} h_{t-1}(\mathbf{x}^{(i)}, y; \theta_{t-1}) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha}\Big]^+ \\ &+ \gamma_t \sum_{i=1}^{N}\sum_{y}\Bigg(\frac{\partial\left[\sum_y h_{t-1}(\mathbf{x}^{(i)}, y; \theta_{t-1}) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha}\right]^+}{\partial\; h_{t-1}(\mathbf{x}^{(i)}, y; \theta_{t-1})} \\ &\qquad \cdot \Big(f_t(\mathbf{x}^{(i)}, y) - h_{t-1}(\mathbf{x}^{(i)}, y)\Big)\Bigg) + O(\gamma_t^2). \end{aligned}$$

Neglecting the second or higher order terms, it is enough to find f_t that minimized the second term,

$$\gamma_t \sum_{i=1}^{N} I\left(\sum_{y} h_{t-1}(\mathbf{x}^{(i)}, y; \theta_{t-1}) c(\mathbf{x}^{(i)}, y) > \tilde{\alpha} \right) \cdot \left(\sum_{y} c(\mathbf{x}^{(i)}, y) f_t(\mathbf{x}^{(i)}, y) \right).$$

As is the case with alternative actions, this term is also minimized by feeding example-dependent cost-sensitive learners with modified training examples $\tilde{E}$, where (15) is modified as

$$\tilde{c}(\mathbf{x}^{(i)}, y) = c(\mathbf{x}^{(i)}, y) \cdot I\left(\sum_{y} h_{t-1}(\mathbf{x}^{(i)}, y; \theta_{t-1}) c(\mathbf{x}^{(i)}, y) > \tilde{\alpha} \right)$$

in the case of allocative actions, since the learners suffer from the cost $c(\mathbf{x}^{(i)}, y)$ only when the argument of I holds.

4. Experiments

In order to compare the risk aversion abilities of cost-sensitive learning and risk-sensitive learning, we conducted two preliminary experiments on a synthetic dataset and a more realistic dataset for credit administration.

4.1 Experimental Settings

First, we explain the implementation and datasets used in the experiments. We used the cost-sensitive perceptron algorithm [6] as the hypothesis $h(\mathbf{x}, y)$ in the case of alternative actions and the weak hypothesis $f_t(\mathbf{x}, y)$ in the case of allocative actions. Especially for the second dataset, we used the kernelized version of the cost-sensitive perceptron with Gaussian kernel to incorporate nonlinearity into the hypothesis. All constant parameters of the perceptron are chosen to have the cost-sensitive perceptron (our baseline method) record the best expected cost with respect to the test data and, they are recycled for the perceptrons used in risk-sensitive learning[†]. This is because we would like to observe the effect of switching the objective function from the cost-sensitive one to the risk-sensitive one. We used the following two datasets.

Synthetic Dataset

In this dataset, there are two-dimensional data $\mathbf{x} = (x_1, x_2)$, and two actions $y \in \{+1, -1\}$. x_1 and x_2 are uniformly randomly sampled over $-5 \leq x_1, x_2 \leq 5$. The cost for each action only depends on x_1 as shown in Figure 4. The cost of action +1 is determined by $c(\mathbf{x}, +1) = 0.1(x_1+5)$ (Figure 4, solid line), and the cost of action −1 is determined by $c(\mathbf{x}, -1) = \mathcal{N}(0, 0.5^2)$ (Figure 4, dashed line). In each experiment, 300 data were generated for training, and 30,000 for test.

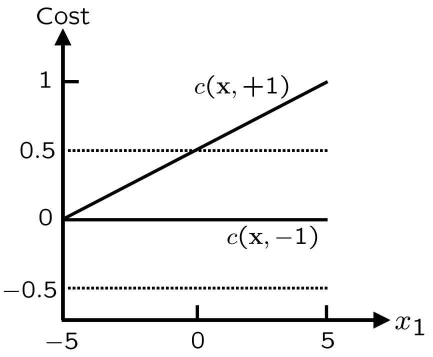

Fig. 4 Expected cost for each action on the synthetic dataset. Note that Gaussian noise $\mathcal{N}(0, 0.5^2)$ with 0.5 standard deviation is added to $c(\mathbf{x}, -1)$ (shown as the dotted lines).

Since the expected cost of action −1 is always smaller that that of action +1, it is enough for cost-sensitive learners to have the trivial hypothesis that always take action −1. However, the costs of action -1 sometimes exceed those of action +1 because of the noise added, In the area of large x_1, it is needed for risk-sensitive learners to switch the action to action +1 to suppress the chance of large costs, since the cost of action +1 is more stable than that of action −1.

Credit Administration

Next, we consider a more realistic application of risk sensitive learning, which is to predict the credit risks of customers. In this task, the learner must predict whether a particular customer can reimburse a loan or not based on his/her profile. Misclassification of a "good customer" as a "bad customer" loses the potential profit, and on the contrary, misclassification of a "bad customer" as a "good customer" loses most of the loan.

We used the "German Credit Data Set" [17] from the STATLOG PROJECT[††] also used in [6]. This dataset includes 700 good customers and 300 bad customers, and $\mathbf{x}$ consists of 20 attributes including sex, age, job, credit history, purpose, and so on. In our experiment, we used the data included in the dataset whose attributes are converted into 24 numerical attributes.

Although the original dataset does not have example-dependent costs, we follow the instruction in [6], and the misclassification cost of a “good customer” as a “bad customer” is defined to be $0.1 \cdot \frac{duration}{12} \cdot$

[†] For example, the width parameter of the Gaussian kernel was determined as $\sigma = 50$.

[††] Data are available from the UCI Machine Learning repository [18].

amount, which means 10% interest per year. The average, variance and maximum cost of this type of cost are 6.27, 43.51^2, and 78.27, respectively. Also, the misclassification cost of a "bad customer" as a "good customer" is defined to be 0.75 × *amount*, which means 75% of the loan is lost. The average, variance and maximum cost of this type of cost are 29.54, 78.09^2, and 138.18, respectively. The other costs are defined to be 0.

While the learner with alternative actions makes binary decisions of whether making loan or not, we can interpret that the learner with allocative actions determines what fraction of the loan is allowed. The realized cost becomes (4) in this case.

4.2 Results

Table 1 and Table 2 show the results for the synthetic data in the cases of alternative actions and allocative actions, respectively. Similarly, Table 3 and Table 4 show the results for the German Credit Data Set. The results for the synthetic data were measured by the averaged values of 5 experiments, and those for the German Credit Data Set were measured by 3-fold cross validation (666 training data and 334 test data). The columns labeled 'Cost-Sensitive' show the results by the cost-sensitive perceptron. The columns labeled 'Risk-Sensitive' show the results by the MetaRisk with $\beta = 0.80, 0.90, 0.95, 0.99$, respectively. Each row shows the values of the CVaR on test data for the corresponding β, and the numbers with ± show the standard errors. The row at the bottom show the mean cost. The values indicated by boldface show the best results among each row.

Overall, as we expected, the cost-sensitive learner has the smallest expected cost, and MetaRisk achieves lower CVaRs than those of the cost-sensitive perceptron at the corresponding βs at the price of the mean cost. When MetaRisk is trained for a particular β, the corresponding test CVaR is almost always better than the test CVaRs by the other training β.

Also, allocative actions achieve better results than alternative actions since the former can realize "portfolios" by combining the costs of two actions. Note that the results for allocative actions are also interpreted as the results from the distribution of the cost expected for each example when the stochastic selection (2) is performed.

Let us examine the cost distributions of the resulted hypotheses. Figure 5 shows the cost distributions for the synthetic data by MetaRisk with alternative actions and $\beta = 0.95$. Even in such a simple case, the cost distribution shows non-Gaussianity since it is a mixture of Gaussian distributions and uniform distribution. Generally, the cost distribution easily becomes non-Gaussian even if each costs follows a Gaussian distributions, since the resulted cost distribution becomes an infinite mixture of Gaussian distributions.

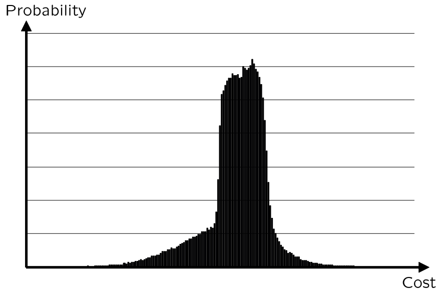

Fig. 5 Cost distribution at $\beta = 0.95$ for the synthetic data shows its non-Gaussianity.

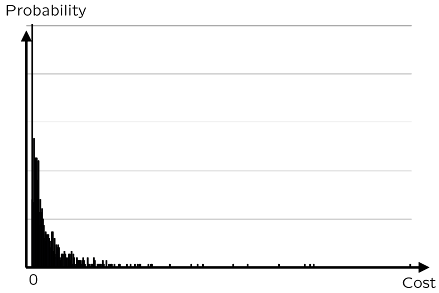

Fig. 6 Cost distribution at $\beta = 0.95$ for the German Credit Data shows its non-Gaussianity.

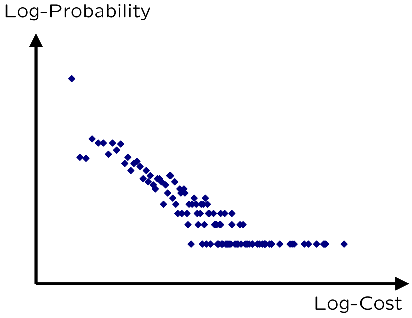

Fig. 7 Double logarithmic plot of cost distribution at $\beta = 0.95$ for the German Credit Data shows its heavy tail property.

Figure 6 the cost distributions for the German Credit Data Set by MetaRisk with alternative actions and $\beta = 0.95$. The cost are significantly skewed to left, and shows its heavy tail property. In order to confirm the heavy tail property, Figure 7 is a double logarithmic plot of the cost distribution. We can observe the linear trend that typical heavy tail distributions show. In both datasets, traditional mean-variance type approaches are not appropriate.

5. Related Work

In this section, we review some works related to risk-sensitive learning, and discuss relations among them.

	Cost-		**Risk-Sensitive**		(β := training β)
test β	**Sensitive**	β = 0.80	β =0.90	β =0.95	β =0.99
β =0.99	1.30±0.01	1.24±0.01	1.19±0.01	1.11±0.05	**1.10**±0.09
β =0.95	0.98±0.01	0.91±0.02	0.84±0.01	**0.83**±0.02	0.98±0.02
β =0.90	0.82±0.01	0.74±0.02	**0.71**±0.01	0.75±0.02	0.93±0.04
β =0.80	0.63±0.01	**0.58**±0.01	0.60±0.01	0.67±0.04	0.85±0.06
Mean Cost	**0.03**±0.01	0.10±0.02	0.17±0.01	0.25±0.04	0.37±0.01

Table 1 Alternative prediction results for synthetic data. The results are shown in the format of 'CVaR ± standard error'.

	Cost-		**Risk-Sensitive**		(β := training β)
test β	**Sensitive**	β = 0.80	β =0.90	β =0.95	β =0.99
β =0.99	1.29±0.00	1.20±0.01	1.12±0.04	1.07±0.06	**0.99**±0.05
β =0.95	0.98±0.00	0.87±0.01	0.81±0.02	**0.80**±0.02	0.87±0.02
β =0.90	0.82±0.01	0.71±0.01	**0.68**±0.01	0.71±0.01	0.82±0.05
β =0.80	0.63±0.01	**0.56**±0.01	0.58±0.01	0.64±0.02	0.75±0.08
Mean Cost	**0.03**±0.01	0.12±0.01	0.20±0.03	0.26±0.05	0.37±0.08

Table 2 Allocative prediction results for synthetic data. The results are shown in the format of 'CVaR ± standard error'.

	Cost-		**Risk-Sensitive**		(β := training β)
test β	**Sensitive**	β = 0.80	β =0.90	β =0.95	β =0.99
β =0.99	64.47±6.46	66.23±6.45	67.91±3.95	60.68±7.06	**55.34**±5.16
β =0.95	34.66±1.14	35.69±1.32	34.45±1.76	**30.13**±2.30	32.00±3.73
β =0.90	23.58±0.94	23.26±0.51	23.04±1.74	**21.15**±1.61	22.89±3.04
β =0.80	14.71±0.78	**14.40**±0.56	14.93±1.16	14.33±1.10	15.38±1.99
Mean Cost	**3.31**±0.23	3.52±1.12	3.92±0.32	3.90±0.34	3.99±0.63

Table 3 Alternative prediction results for the German Credit Data Set [17] (3-fold cross validation). The results are shown in the format of 'CVaR ± standard error'.

	Cost-		**Risk-Sensitive**		(β := training β)
test β	**Sensitive**	β = 0.80	β =0.90	β =0.95	β =0.99
β =0.99	64.47±6.46	60.29±6.70	57.89±3.71	52.74±1.61	**44.29**±7.28
β =0.95	34.66±1.14	31.48±2.03	30.03±1.45	**26.17**±1.27	28.16±3.19
β =0.90	23.58±0.94	20.76±1.36	20.25±1.25	**19.21**±0.85	22.45±2.31
β =0.80	14.71±0.78	**13.01**±0.80	13.73±0.70	14.47±0.58	16.65±1.87
Mean Cost	**3.31**±0.2	3.98±0.13	4.68±0.22	5.08±0.17	5.55±0.51

Table 4 Allocative prediction results for the German Credit Data Set [17] (3-fold cross validation). The results are shown in the format of 'CVaR ± standard error'.

5.1 Financial Engineering

Decision making theory considering risk aversion originates [19]'s mean-variance model, and thereafter, has been actively studied as the portfolio theory in the fields of operations research and financial engineering [9]. Value-at-Risk (VaR) [12] is probably the most commonly used metric of risks. The convexity of the optimization problem of VaR is guaranteed if the underlying cost distribution follows Gaussian distribution, but this assumption does not hold in many real situations. Recently, a new risk metric called conditional value-at-risk (CVaR) [10] (a.k.a. expected shortfall) is attracting considerable attention since it considers the amount of costs exceeding VaR, and conveniently, is convex without the Gaussian assumption of the cost distribution. Most of the works in this field focus on estimating the amount of risks [12], or solving mathematical programming for optimal decision making given models [13], and there are little works [22]–[24] from machine learning perspective such as learning risk-avoiding decision rules from examples.

5.2 Cost-Sensitive Learning

There are many types of costs treated in cost-sensitive learning [20]. In early works [1]–[3], the costs are assumed to be known, and not to depend directly on $\mathbf{x}$, but on classes as latent variables. Recently, direct minimization of the expected cost (6) in more general situations where the costs are not known beforehand, and depend on $\mathbf{x}$, has been widely accepted [4]–[8]. There are three types of approaches in existing cost-sensitive learners. One is decision-theoretic approaches that perform Bayes-optimal decision making based on estimated class probabilities and cost distributions [1], [5]. Another approach is the cost-sensitive versions of the existing cost-insensitive learners such as decision trees [2], perceptrons [6], and support vector machines [6], [21], and the other approach is meta-learners that exploit existing cost-insensitive learners to realize cost-sensitive learning by reweighting or resampling examples [3], [4], [6]–[8]. However, all works are oriented toward minimizing the expected cost, and not toward mitigating the risks of huge costs as we discussed in this paper.

5.3 Risk-Sensitive Reinforcement Learning

After the seminal work by [22], there are some attempts to incorporate the idea of risk aversion in the context of reinforcement learning [23], [24]. However, they all remain to focus on minimizing the value-at-risk in limited cases. For example, instead of the expected discounted reward, Herger proposes the α-value criterion [22] as the objective function, which is essentially identical to value-at-risk of the discounted reward, and which is not convex. Also, the Bellman equation is presented for the worst case, that is, $\beta = 1$, and it is not possible for general β. [23] realized soft risk aversion by employing a parameter that emphasizes actions whose rewards are less than expected, but this parameter is rather heuristic, and does not have clear correspondence to the risk metric to be optimized.

The above methods are both designed as variants of Q-learning, hence do not aim directly to optimize the risk metrics, but aim to estimate the expected discounted reward function accurately. On the other hand, [24] propose an approach that directly optimizes an objective function defined as a linear combination of the mean and the variance of discounted reward. This is based on the assumption of the mean-variance model where the distribution of the discounted reward follows Gaussian distribution, which does not hold in most situations. Moreover, in the case of alternative actions, the objective function is not convex even under the assumption.

5.4 Robust Statistics

Robust statistics [25], [26] aims to robust estimation of models by eliminating influence of outliers, which is antithetical to our risk-sensitive approach. There are some classes of robust estimators, one of which is the L-estimator defined as a linear combination of order statistics. For example, least trimmed square [26] is an instance of the L-estimator which minimizes the sum of squared losses less than some quantile by trimming off the largest losses. The idea of trimmed estimator has been generalized for general loss functions [27].

Since minimization of CVaR is identical to minimization of the losses above β-quantile, our approach goes against the trimmed estimators in that sense. In contrast to cutting off outliers to robustify estimators, it makes the most of the outliers, and aggressively "overfits" to them to avoid potential risks. In addition, trimmed estimation is not usually a convex optimization problem, while minimizing CVaR is.

6. Conclusion and Future Works

In this paper, we tackled cost-sensitive learning problem from the perspective of risk aversion, and proposed to minimize not the expected cost but the risk metric called conditional value-at-risk (CVaR) which is being widely accepted in the area of financial engineering. Its definition and characteristics such as convexity play key roles to elegantly realize risk aversion, which has not been discussed in the area of data mining. The proposed method is a meta-learning algorithm that exploits any existing cost-sensitive learner to solve risk-sensitive learning problems with alternative or allocative actions.

Although we focused on supervised classification problems in this paper, this idea is also applicable to a wide class of data mining problems, such as clustering, regression, and so on. Also, from another perspective, the cost can be substituted by general loss functions such as log-likelihood. This indicates that the meta-learning framework we proposed in this paper has possibilities of converting existing machine learning algorithms to have large margin and sparsity properties by enforcing them to focus on difficult examples just like boosting and suport vector machines.

Finally, we conclude this paper with mentioning some possible future works. Although we used the CVaR in a stand-alone manner in this paper, there might be cases where one wants not only to minimize risks of large costs but also to minimize the expected cost at the same time. Actually, such idea is widely accepted in portfolio theory that maximizes expected returns while suppressing risks. Similarly, we should incorporate the expected cost into the objective function in real applications. One way to do this is to employ a linear combination of the expected cost (6) and the risk metric (9) as the objective function,

$$\eta C^E(\theta) + (1 - \eta)\phi^D_\beta(\theta),$$

where $0 \leq \eta \leq$ is a mixing constant. It is easily confirmed that this objective function also has convexity, and MetaRisk can be extended to afford this objective function.

Another possibility is to develop tailor-made algorithms for risk-sensitive learning that minimize (12) with respect to both θ and α at the same time, while MetaRisk optimizes θ and α alternately in this paper. In the case of allocative actions, we assumed a linear constraint (4) on the cost of action portfolio, and this made it possible to take the gradient boosting approach. However, this assumption might be too strong in some applications. As Theorem 1 assures, CVaR is convex if we make a more general assumption that $c(\mathbf{x}, h(\theta))$ is convex. Approaches from direct convex optimization might be pursued in such cases. From the viewpoint of computational efficiency, perceptron learning that we employed in the experiment is incremental with respect to θ, but MetaRisk itself is a batch algorithm. This is not efficient for large data sets, and thoroughly on-line type algorithms are desirable.

The other direction of the future research is to loosen the assumption on the training data. The assumption that we know costs for all actions seems to be too strong. There should be many cases where we know the cost for the action we really took, for example, data on direct marketing usually has the results only for the actions that were actually taken. These kinds of situations might be modeled as a one-benefit learning problem [28], or similarly, an associative reinforcement learning problem [29], [30]. More generally, reinforcement learning with the CVaR of discounted reward might be seen beyond these problems.

Acknowledgement

The author thanks Rikiya Takahashi, Tsuyoshi Idè, Takayuki Osogami, Yuta Tsuboi and Ryo Sugihara of IBM Tokyo Research Laboratory for fruitful discussion. The author also thanks Koji Tsuda of MPI for Biological Cybernetics for suggestive comments.

References

[1] C. Elkan, "The foundations of cost-sensitive learning," Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), pp.973–978, 2001.

[2] J.P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C.E. Brodley, "Pruning decision trees with misclassification costs," Proceedings of the 9th European Conference on Machine Learning (ECML), 1998.

[3] P. Domingos, "MetaCost: A general method for making classifier cost sensitive," Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, pp.155–164, 1999.

[4] W. Fan, S.J. Stolfo, J. Zhang, and P.K. Chan, "AdaCost: Missclassification cost sensitive boosting," Proceedings of the 16th International Conference on Machine Learning (ICML), pp.97–105, 1999.

[5] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," Proceedings of ACM SIGKDD Conference, 2001.

[6] P. Geibel, U. Bredford, and F. Wysotzki, "Perceptron and SVM learning with generalized cost models," Intelligent Data Analysis, vol.8, no.5, pp.439–455, 2004.

[7] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," Proceedings of the 3rd International Conference on Data Mining (ICDM), pp.435–442, 2003.

[8] N. Abe and B. Zadrozny, "An iterative method for multi-class cost-sensitive learning," Proceedings of ACM SIGKDD Conference, 2004.

[9] D.G. Luenberger, Investment Science, Oxford University Press, 1998.

[10] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath, "Coherent measures of risk," Mathematical Finance, vol.9, pp.203–228, 1999.

[11] W. Hoeffding, "Probability inequalities for sums of bounded random variables," Journal of the American Statistical Association, vol.58, pp.13–30, 1963.

[12] H. Mausser and D. Rosen, "Beyond VaR: From measuring risk to managing risk," ALGO Research Quarterly, vol.1, no.2, pp.5–20, 1998.

[13] R.T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," Journal of Risk, vol.2, no.3, pp.21–41, 2000.

[14] T.C. Cormen, C.E. Leiserson, and R.L. Rivest, Introduction to Algorithms, MIT Press, Cambridge, MA, 1990.

[15] J.H. Friedman, "Greegy function approximation: A gradient boosting machine," Annals of Statistics, vol.29, no.5, 2001.

[16] S. Rosset and E. Segal, "Boosting density estimation," Advances in Neural Information Processing Systems 15, 2002.

[17] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, Machine Learning, Neural and Statistical Classification, Ellis Horwood, 1994.

[18] D.J. Newman, S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998.

[19] H.M. Markovitz, "Portfolio selection," Journal of Finance, vol.7, no.1, pp.77–91, 1952.

[20] P. Turney, "Types of cost in inductive concept learning," Proceedings of Workshop on Cost-Sensitive Learning (WSCL) at the 17th International Conference on Machine Learning (ICML), pp.15–21, 2000.

[21] G. Fumera and F. Roli, "Cost-sensitive learning in support vector machines," VIII Convegno Associazione Italiana per L'Intelligenza Artificiale, 2002.

[22] M. Herger, "Considering of risk in reinforcement learning," Proceedings of the 11th International Conference on Machine Learning (ICML), pp.105–111, 1994.

[23] R. Neuneier, "Risk sensitive reinforcement learning," Advances in Neural Information Processing Systems 11, 1998.

[24] M. Sato, H. Kimura, and S. Kobayashi, "TD algorithm for the variance of return and mean-variance reinforcement learning [in Japanese]," Journal of Japanese Society of Artificial Intelligence, vol.16, no.3, pp.353–362, 2001.

[25] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, John Wiley and Sons, New York, 1986.

[26] P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, John Wiley and Sons, New York, 1987.

[27] P. Cizek, "General trimmed estimation: Robust approach to nonlinear and limited dependent models," Tech. Rep. 2004-1300, CentER Discussion Paper, 2004.

[28] B. Zadrozny, "One-benefit learning: cost-sensitive learning with restricted cost information," Proceedings of the 1st International Workshop on Utility-based Data Mining (UDBM), pp.53–58, 2005.

[29] L.P. Kaelbling, "Associative reinforcement learning: Function in k-DNF," Machine Learning, vol.15, no.3, pp.279–298, 1994.

[30] R.J. Williams, "Simple statistical gradient following algorithms for connectionist reinforcement learning," Machine Learning, vol.8, no.3, pp.229–256, 1992.

Hisashi Kashima is a researcher of Tokyo Research Laboratory of IBM Research since 1999. He received M.E. and Ph.D. degrees from Kyoto University in 1999 and 2007, respectively, His research interests include machine learning and its application to bioinformatics, autonomic computing and business intelligence.