Risk-Sensitive Learning via Expected Shortfall Minimization

Hisashi Kashima*

Abstract

A new approach for cost-sensitive classification is proposed. We extend the framework of cost-sensitive learning 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 expected shortfall, a.k.a. conditional value-at-risk, known as a good risk metric in the area of financial engineering. The proposed algorithm is a general meta-learning algorithm that can utilize 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.

Keywords

Risk-Sensitive Learning, Cost-Sensitive Learning, Meta Learning, Risk Management, Expected Shortfall, Conditional Value-at-Risk

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. Commonly, classification algorithms are designed 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.

Cost-sensitive learning [4, 3, 10, 5, 11, 1] is a suitable framework for such cases where costs are different among data, 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 with 0-1 costs. However, from the standpoint of risk management, there

are situations where cost-sensitive learning is still not enough. If there is not a little chance of huge cost occurring, and also if users are interested in mitigating the risk, it can not avoid such a risk of disasters since it does not aggressively suppress the occurrence of huge costs. Risk aversion is one of the central topics in financial engineering. For example in portfolio theory, it is expected to find a portfolio that maximizes profit while suppressing the risks of huge costs occurring with low probabilities [7].

In this paper, we propose an approach of risk-sensitive classification that considers cost distributions not to decrease the expected cost, but to mitigate the risks of huge costs. Instead of the expected cost, we employ a risk metric called expected shortfall [2], a.k.a. conditional value-at-risk. We propose a risk-sensitive learning algorithm that minimizes the expected shortfall as the objective function. Also, our algorithm is a meta-learning algorithm, which is quite a general procedure that can convert existing cost-sensitive learners to risk-sensitive learners.

2 Drawback of Cost-Sensitive Learning

We first review cost-sensitive learning with example-dependent costs, and point out its drawback from the standpoint of risk management.

Let X be a set of all *target objects*, e.g. $X = \mathbb{R}^M$, 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 $h(\mathbf{x}, y; \boldsymbol{\theta}) : X \times Y \to \mathbb{R}$, where $\boldsymbol{\theta}$ is its model parameters. An action $\hat{y} \in Y$ taken against $\mathbf{x} \in X$ is determined by

$$(2.1) \qquad \hat{y} = \operatorname*{argmax}_{y \in Y} h(\mathbf{x}, y; \boldsymbol{\theta}).$$

Usually, only one action is assumed to be taken at a time, hence we call this type of actions *alternative actions*. If it is allowed to take multiple actions at a time, and to allocate resources to each of $|Y|$ actions in proportion to $h(\mathbf{x}, y; \boldsymbol{\theta})$ with the following constraint,

$$(2.2) \qquad \sum_{y \in Y} h(\mathbf{x}, y; \boldsymbol{\theta}) = 1, \text{ s.t. } h(\mathbf{x}, y; \boldsymbol{\theta}) \geq 0,$$

*Tokyo Research Laboratory, IBM Research

for $\forall \mathbf{x} \in X, \forall y \in Y$, those kind of actions are called *allocative*. Allocative actions are popular in the context of portfolio selection [7] 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 (2.1), and in the other of which resources are allocated with (2.2).

Cost function is a function $c(\mathbf{x}, y) : X \times Y \rightarrow \mathbb{R}$, which indicates how bad an action $y \in Y$ taken against $\mathbf{x} \in X$ is. For instance in medical diagnosis, $c(\mathbf{x}, y)$ becomes small if the treatment is appropriate, and becomes large if not. We deal with the most general problem setting in cost-sensitive learning, where the true cost function is unknown, and depends on examples [10, 5, 1]. Let $c(\mathbf{x}, h(\boldsymbol{\theta}))$ be the cost of the action for $\mathbf{x}$ by using hypothesis $h(\mathbf{x}, y; \boldsymbol{\theta})$. In the case of alternative actions (2.1), $c(\mathbf{x}, h(\boldsymbol{\theta}))$ becomes

$$(2.3) \qquad c(\mathbf{x}, h(\boldsymbol{\theta})) = c(\mathbf{x}, \operatorname*{argmax}_{y \in Y} h(\mathbf{x}, y; \boldsymbol{\theta})).$$

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

$$(2.4) \qquad c(\mathbf{x}, h(\boldsymbol{\theta})) = \sum_{y \in Y} h(\mathbf{x}, y; \boldsymbol{\theta}) c(\mathbf{x}, y),$$

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 [7].

Cost-sensitive learning [4, 3, 10, 5, 11, 1] is a framework for supervised classification learning with cost functions $c(\mathbf{x}, y)$. The expected cost is conventionally used as the objective function for training to find the best $\boldsymbol{\theta}$. The expected cost with respect to data distribution D over $X \times \mathbb{R}^Y$ is defined as

$$(2.5) \qquad C^D(\boldsymbol{\theta}) = E_D[c(\mathbf{x}, h(\boldsymbol{\theta}))].$$

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 parameter $\boldsymbol{\theta}$ is determined so that the following empirical expected cost $C^E(\boldsymbol{\theta})$ is minimized [10, 5, 1],

$$(2.6) \qquad C^E(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})).$$

However, let us imagine such a situation where the occurrences of huge costs are fatal. For example, if we would like to decide where to invest our fund, several consecutive failed investments might directly leads to risk of bankruptcy. In such 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. Let us consider another example. Assume that two hypotheses $h(\boldsymbol{\theta}_1)$ and $h(\boldsymbol{\theta}_2)$, and both of them have identical expected costs. $h(\boldsymbol{\theta}_1)$ has a cost distribution with high peak around its expected cost, and $h(\boldsymbol{\theta}_2)$ has one with a gentle slope and a heavy tail in its high cost area. In this situation, risk aversive investers would apparently prefer $h(\boldsymbol{\theta}_1)$ to $h(\boldsymbol{\theta}_2)$. The above discussion implies us that minimization of the expectation of $c(\mathbf{x}, h(\boldsymbol{\theta}))$ is not enough, and suggests the need to consider the distribution of $c(\mathbf{x}, h(\boldsymbol{\theta}))$ and aggressively avoid the risk of huge costs.

3 Risk-Sensitive Learning via Expected Shortfall Minimization

Motivated by the discussion in the previous section, we propose our risk-sensitive learning approach using a new objective function that aggressively avoids the risk of huge costs, and then propose a meta-learning algorithm that reduces cost-senstive learners to risk-sensitive learners.

3.1 Expected Shortfall *Expected shortfall* [2], a.k.a. conditional value-at-risk, 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 for a given constant $0 \leq \beta \leq 1$ (See Figure 1.), hence it can consider the amount of huge costs. Moreover, expected shortfall has desirable characteristics such as convexity [8].

In our setting, the expected shortfall $\phi_\beta^D(\boldsymbol{\theta})$ with respect to hypothesis h and data distribution D is defined as

(3.7)

$$\phi_\beta^D(\boldsymbol{\theta}) = \frac{1}{1-\beta} E_D\big[I\big(c(\mathbf{x}, h(\boldsymbol{\theta})) \geq \alpha_\beta^D(\boldsymbol{\theta})\big)c(\mathbf{x}, h(\boldsymbol{\theta}))\big]$$

$$\alpha_\beta^D(\boldsymbol{\theta}) = \min\big\{\alpha \in \mathbb{R} \mid E_D\big[I\big(c(\mathbf{x}, h(\boldsymbol{\theta})) \geq \alpha\big)\big] \leq 1 - \beta\big\},$$

where $\alpha_\beta^D(\boldsymbol{\theta})$ is called *value-at-risk* (VaR) [2], i.e. the β-quantile of cost distribution, and $I(\cdot)$ is a function that returns 1 when its argument is true, and returns 0 otherwise. Since the expected shortfall is the expected costs surpassing $\alpha_\beta^D(\boldsymbol{\theta})$, (3.7) is decomposed into two terms as
(3.8)

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

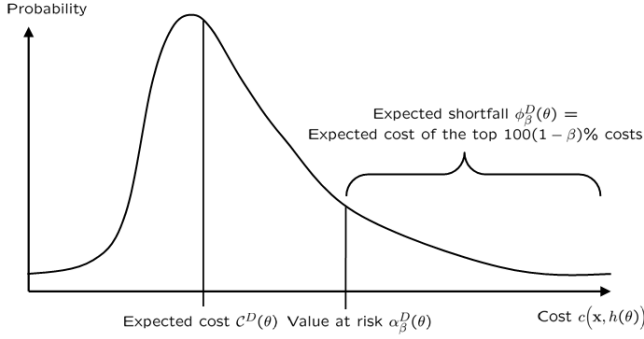where $[x]^+$ is a function that returns x when $x \geq 0$, and

Figure 1: Expected shortfall.

Figure 2: MetaRisk: A risk-sensitive meta-learner.

returns 0 otherwise.

3.2 MataRisk: A Risk-Sensitive Learner to Minimize Expected Shortfall

Let us derive an algorithm to optimize parameter $\boldsymbol{\theta}$. Instead of (3.8), we employ the following empirical expected shortfall defined on training examples E instead of D which is unknown.

(3.9)
$$\phi_\beta^E(\boldsymbol{\theta}) = \alpha_\beta^E(\boldsymbol{\theta}) + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} \left[c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) - \alpha_\beta^E(\boldsymbol{\theta}) \right]^+,$$

where $\alpha_\beta^E(\boldsymbol{\theta})$ is the value-at-risk for the training examples E,
(3.10)
$$\alpha_\beta^E(\boldsymbol{\theta}) = \min \left\{ \alpha \in \mathbb{R} \, \big| \, \frac{1}{N} \sum_{i=1}^{N} I\big(c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) \geq \alpha\big) \leq 1 - \beta \right\}.$$

Now, if we suppose that $\alpha_\beta^E(\boldsymbol{\theta})$ is a known constant $\tilde{\alpha}$ in (3.9), we only have to minimize the second term of the second term (3.9),

(3.11)
$$\tilde{C}_{\tilde{\alpha}}^E(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} \left[c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) - \tilde{\alpha} \right]^+.$$

Note that (3.11) is convex if $c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ is convex with respect to $\boldsymbol{\theta}$. For the time being, we assume existence of algorithms to find $\boldsymbol{\theta}$ that minimizes (3.11). Next, we fix $\boldsymbol{\theta}$, and find the VaR (3.10) for the $\boldsymbol{\theta}$. This is equivalent to finding $c(\mathbf{x}^{(k)}, h(\boldsymbol{\theta}))$ where k is the index of the training datum with the $\lfloor (1 - \beta)N \rfloor$-th largest cost by $\boldsymbol{\theta}$. Based on the above discussion, we propose a risk-sensitive meta-learning algorithm named MetaRisk (Figure 2)[1], which minimizes the empirical expected shortfall by exploiting existing cost-sensitive

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

learners, and by finding the model parameter and the corresponding value-at-risk alternately.

The optimality and convergence of MetaRisk are directly guaranteed by the following theorem that shows the convexity of the upper bound of expected shortfall.

THEOREM 3.1. ([8], THEOREM 1&2) *Let*
(3.12)
$$F_\beta^E(\boldsymbol{\theta}, \alpha) = \alpha + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} \left[c(\boldsymbol{x}^{(i)}, h(\boldsymbol{\theta})) - \alpha \right]^+,$$

then

(3.13)
$$\min_{\boldsymbol{\theta}} \phi_\beta^E(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}, \alpha} F_\beta^E(\boldsymbol{\theta}, \alpha).$$

$F_\beta^E(\boldsymbol{\theta}, \alpha)$ *is convex with respect to* α. *If* (2.6) *is convex with respect ot* $\boldsymbol{\theta}$, $F_\beta^E(\boldsymbol{\theta}, \alpha)$ *is also jointly convex with respect to* $\boldsymbol{\theta}$ *and* α. *Also,*

(3.14)
$$\alpha_\beta^E(\boldsymbol{\theta}) = \min\{\alpha \in \operatorname*{argmin}_{\alpha} F_\beta^E(\boldsymbol{\theta}, \alpha)\}$$

holds. □

(3.13) indicates that minimization of (3.12) is equivalent to minimization of expected shortfall. The joint convexity of (3.12) ensures the gradient-based optimization with respect to $\boldsymbol{\theta}$ and α. Moreover, from (3.14), $\alpha_\beta^E(\boldsymbol{\theta})$ is the minimizer of $F_\beta^E(\boldsymbol{\theta}, \alpha)$ at $\boldsymbol{\theta}$, hence MetaRisk exactly performs coordinate-wise descent of $F_\beta^E(\boldsymbol{\theta}, \alpha)$.

3.3 Recycling Existing Cost-Sentive Learners

We propose methods to minimize (3.11) by calling existing cost-sensitive learners with reweighted costs.

Reduction is relatively easy in the case of alternative actions (2.1). Paying attentions to its similarity to (2.6), we notice that this is the expectation of only costs exceeding $\alpha_\beta^E(\boldsymbol{\theta})$. 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

(3.15)
$$\tilde{c}^{(i)}(\mathbf{x}^{(i)}, y) = [c^{(i)}(\mathbf{x}^{(i)}, y) - \tilde{\alpha}]^+$$

for the original costs, (2.6) becomes

529

$$(3.16) \qquad \tilde{C}_{\tilde{\alpha}}^{E}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \tilde{c}^{(i)}(\mathbf{x}^{(i)}, y),$$

and this has the same form as the expected cost (2.6). The reduction is realized by feeding example-dependent cost-sensitive learners [5, 11, 1] with modified training examples $\tilde{E}$, where the i-th example of $\tilde{E}$ is defined as $\tilde{e}^{(i)} = \left(\mathbf{x}^{(i)}, \{\bar{c}^{(i)}(\mathbf{x}^{(i)}, y)\}_{y \in Y} \right)$.

Next, let us consider the case where stochastic or allocative decision making by the constrained hypothesis (2.2) is allowed. (3.11) is rewritten as

$$(3.17) \quad \tilde{C}_{\tilde{\alpha}}^{E}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \Big[\sum_{y} h(\mathbf{x}^{(i)}, y; \boldsymbol{\theta}) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha} \Big]^{+}.$$

Unlike the case of alternative actions, $c^{(i)}(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ depends on a convex combination of $c^{(i)}(\mathbf{x}^{(i)}, y)$, hence simple reweighting like (3.15) does not work. Although a natural choice of such a classifiers is the exponential family satisfying (2.2) such as multi-class logistic regression, $c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ is not convex with respect to its parameters, and even worse, it can be a multi-modal function. Therefore, we employ a family of classifiers with which $c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ is linear with respect to $\boldsymbol{\theta}$. (3.17) is convex with respect to its parameters. Especially, we employ the gradient boosting method [9], where $h(\mathbf{x}, y; \boldsymbol{\theta})$ is represented as a linear combination of T deterministic hypotheses $f_1, \ldots, f_T$,

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

where $\boldsymbol{\theta}_t = (w_1, \ldots, w_t)$ are the parameters. Since $h(\mathbf{x}, y; \boldsymbol{\theta})$ has to satisfy the stochastic constraints (2.2), we need $\sum_{t=1}^{T} w_t = 1$, 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

$$h_t(\mathbf{x}, y; \boldsymbol{\theta}_t) = (1 - \gamma_t) h_{t-1}(\mathbf{x}, y; \boldsymbol{\theta}_{t-1}) + \gamma_t f_t(\mathbf{x}, y),$$

where $0 < \gamma_t \leq 1$ is a updating parameter at round t, and $w_t = \gamma_t \prod_{\tau = t+1}^{T} (1 - \gamma_\tau)$.

In order to find f_t at round t, assume that γ_t is sufficiently small, then the second order term of the Taylor series expansion of (3.17) around h_{t-1} gives

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

As is the case with alternative actions, this is also minimized by feeding example-dependent cost-sensitive learners with modified training examples $\tilde{E}$, where (3.15) is modified as

$$\tilde{c}(\mathbf{x}^{(i)}, y) = c(\mathbf{x}^{(i)}, y) I\Big(\sum_{y} h_{t-1}(\mathbf{x}^{(i)}, y; \boldsymbol{\theta}_{t-1}) c(\mathbf{x}^{(i)}, y) > \tilde{\alpha} \Big)$$

in the case of allocative actions.

4 Experiments

We conducted a preliminary experiment on a dataset for credit administration. In this task, the learner must predict whether a particular customer can make a loan or not based on his/her profile. Misclassification of a "good customer" as a "bad customer" loses the potential interest, 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" from the STATLOG PROJECT[2] also used in [5]. This dataset includes 700 good customers and 300 bad customers, and $\mathbf{x}$ consists of 24 attributes including sex, age, job, credit history, purpose, and so on. Although the original dataset does not have example-dependent costs, we follow the instruction in [5], and the misclassification cost of a "good customer" as a "bad customer" is defined to be $0.1 \cdot \frac{duration}{12} \cdot amount$, 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 $0.75 \times amount$, 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 (2.4) in this case.

We used the kernelized version of the cost-sensitive perceptron algorithm [5] with Gaussian kernel[3] as the cost-sensitive learner. Table 1 show the results in the cases of alternative actions and allocative actions 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 expected shortfall on test data for the corresponding β, and the numbers in the brackets show the value-at-risks. The row at the

---

[2]Data are available from UCI Machine Learning repository [6].
[3]The width parameter of the Gaussian kernel was determined as $\sigma = 50$ so that the cost-sensitive perceptron record the best expected cost.

Table 1: Experimental results for alternative actions (top) and allocative actions (bottom).

Test ES	Cost-	Risk-Sensitive			
(VaR)	Sensitive	$\beta = 0.80$	$\beta = 0.90$	$\beta = 0.95$	$\beta = 0.99$
$\beta = 0.99$	64.47 (46.59)	66.23 (50.66)	67.91 (50.92)	60.68 (38.21)	**55.34** (40.66)
$\beta = 0.95$	34.66 (17.67)	35.69 (15.28)	34.45 (14.74)	**30.13** (15.28)	32.00 (17.96)
$\beta = 0.90$	23.58 (10.16)	23.26 (7.58)	23.04 (9.09)	**21.15** (10.45)	22.89 (11.58)
$\beta = 0.80$	14.71 (3.30)	**14.40** (3.66)	14.93 (5.12)	14.33 (5.48)	15.38 (5.03)
Mean Cost	**3.31**	3.52	3.92	3.90	3.99

Test ES	Cost-	Risk-Sensitive			
(VaR)	Sensitive	$\beta = 0.80$	$\beta = 0.90$	$\beta = 0.95$	$\beta = 0.99$
$\beta = 0.99$	64.47 (46.59)	60.29 (45.53)	57.89 (41.25)	52.74 (38.47)	**44.29** (30.69)
$\beta = 0.95$	34.66 (17.67)	31.48 (14.06)	30.03 (13.26)	**26.17** (14.13)	28.16 (20.65)
$\beta = 0.90$	23.58 (10.16)	20.76 (7.42)	20.25 (8.95)	**19.21** (11.31)	22.45 (13.79)
$\beta = 0.80$	14.71 (3.30)	**13.01** (3.74)	13.73 (6.09)	14.47 (8.20)	16.65 (8.77)
Mean Cost	**3.31**	3.98	4.68	5.08	5.55

bottom show the mean cost. The values indicated by boldface show the best results among each row. The results show that MetaRisk achieves lower expected shortfalls than those of the cost-sensitive perceptron at the corresponding βs at the price of the mean cost. Allocative actions achieve better results than alternative actions since the former can realize "portfolios" by combining the costs of two actions.

5 Conclusion

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 expected shortfall. 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 problems with alternative or allocative actions.

Although we focused on supervised classification in this paper, the basic idea is also applicable to a wide class of data mining problems, such as clustering, regression, and so on. Also, 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 SVM.

References

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

[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

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

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

[5] P. Geibel, U. Bredford, and F. Wysotzki. Perceptron and SVM learning with generalized cost models. *Intelligent Data Analysis*, 8(5):439–455, 2004.

[6] S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998.

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

[8] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

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

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

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