2005 年情報論的学習理論ワークショップ
2005 Workshop on Information-Based Induction Sciences (IBIS2005)
Tokyo, Japan, November 9 - 11, 2005.

リスク回避型学習
Risk-Sensitive Classification Learning

鹿島 久嗣*
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. Experiments on tasks with example-dependent costs show promising results.
keywords: Cost-Sensitive Learning, Risk Management, Financial Engineering, Expected Shortfall, Meta Learning

1 はじめに

分類学習は, データマイニングにおける基本的なタスクの 1 つであり, 対象とクラス（あるいはその対象に対してとるべきアクション）の組が, 訓練データとして複数与えられたときに, クラス未知の対象に対し, その対象が所属するクラス（アクション）を正しく予測すること, 言い換えれば, 誤り回数（誤り確率）を最小化することを目的としている. この問題は, 例えば, 医療分野における診断や, 金融分野における与信審査, 検索エンジンにおける Web ページのトピック分類などのように, 世の中のいたるところで現れる一般的かつ重要な問題である. しかしながら, 例えば, ある患者を, 本当は健康であるのにもかかわらず病気だと判断してしまい余計な治療を行ってしまう誤りのコストと, 本当は病気であるのにもかかわらず, 放置してしまい, 深刻な結果を引き起こしてしまう誤りのコストは, 明らかに対称であるとはいい難く, また, その大きさも患者によって異なってくるであろう.

単に誤り回数を最小化するだけでは十分ではなく, 対象ごとに, あるいはクラスごとに誤りコストが異なり, またそれらが予測時に未知であるような場合を許容する枠組みとして, コスト考慮型 (Cost-Sensitive) 学習 [18, 6, 3, 5, 7, 14, 9, 15, 1] がある. コスト考慮型学習では, 単にクラス未知の対象に対する誤り率を最小化するのではなく, 期待コストを最小化することを目的としているため, より広範囲の問題を扱えるようになる[1]. しかしながら, リスクマネジメントの立場からは, 期待コストを最小化するコスト考慮型学習では十分でない状況が存在する. 期待コストの最小化は全体としては発生するコストを軽減するであろうが, これは, 必ずしも高いコストの発生するリスクを積極的に回避することを目指してはいるわけではないため, 対象によっては思いがけず大きなコストが生じる場合を抑えることができない. 起こる確率は小さいが, 大きなコストが発生してしまうような可能性があり, ユーザーがそのリスクをなるべく回避することに興味がある場合, 期待コストではこの目的を正しく反映しているとはいえない. 特に, 金融工学においてはリスク回避は主要なテーマとなっており, 例えば, 株式投資の場合には, 小さな確率でおきる大きな損失を回避しながら収益を高めるようなポートフォリオを組むことが必要となる [17].

本論文では, コストの分布を考慮し, 期待コストではなく, 大きなコストが発生するリスクを抑えるような, リスク回避型 (Risk-Sensitive) 分類学習手法を提案する. 具体的には, 期待コストの代わりに, 金融分野において資産のもつリスクを測る指標として注目されている期待ショートフォール（条件付きバリュー・アット・リスク）[16, 2] と呼ばれる量を目的関数として用いる学習アルゴリズムを提案する. このアルゴリズムは, メタ学習アルゴリズムであり, 既存のコスト考慮型学習アルゴリズム

*日本アイ・ビー・エム株式会社 東京基礎研究所, 242-8502 神奈川県大和市下鶴間 1623-14, tel. 046-215-4698, e-mail hkashima@jp.ibm.com,
Tokyo Research Laboratory, IBM Research, 1623-14 Shimotsuruma, Yamato-shi, 242-8502 Kanagawa, Japan

[1] 通常の分類問題は, すべての誤りのコストが等しく 1 であると仮定していると解釈できる.

を, リスク回避型学習アルゴリズムに変換することのできる汎用的な手法になっている. また, 本論文で我々が示すアプローチは, 教師あり分類問題に限定することなく, 様々な学習モデルにおいても適用可能であり, この意味でも非常に一般的であるといえる.

2 コストを考慮した学習

この章では, まず, 従来のコストを考慮する学習の基本とその問題点を述べる.

2.1 意思決定モデル

X を対象の集合 (たとえば $X = \mathbb{R}^M$) とし, Y をこれらに対してとりうるアクションの集合（離散的で有限）とする. 対象 $\mathbf{x} \in X$ に対し, アクション $y \in Y$ を決定するために用いる関数を

$$h(\mathbf{x}, y; \boldsymbol{\theta}) : X \times Y \to \mathbb{R}$$

とする. ここで, $\boldsymbol{\theta}$ はモデルのパラメータである. これを用いて通常, 取るべきアクション $\hat{y}$ は,

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

によって択一的に決定される.

$h(\mathbf{x}, y; \boldsymbol{\theta})$ には以下のように確率的な制約が入っていてもよい.

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

対象 $\mathbf{x} \in X$ が与えられたとき, これに対してこの場合, アクションの決定は (1) の代わりに, (2) によって確率的に決定してもよい.

また, アクションの決定が資源分散型, すなわち, 実際にとることのできるアクションがひとつではなく $h(\mathbf{x}, y; \boldsymbol{\theta})$ の割合に応じて資源を分散投資できるような場合も考えられる. この場合については後に議論するが, ひとまずは, アクションは (1) によって択一的に決定されるものとして議論を進める.

2.2 コスト関数

コストとは, ある対象にとってとったアクションに対して, それがどのくらい悪かったかをあらわす指標として与えられる[2]. 対象 $\mathbf{x} \in X$ に対し, アクション $y \in Y$ をとったときのコストを $c(\mathbf{x}, y) \in \mathbb{R}$ とおく. 例えば, ある検査結果 $\mathbf{x}$ であった患者に, ある治療 y を行なったときに引き起こされる結果の悪さの程度が $c(\mathbf{x}, y)$ となる. 正しい治療なら $c(\mathbf{x}, y)$ は小さく, 誤った治療なら大きくなる. この患者に対する治療として y が非常に不適切で, より健康状態を損なってしまった場合, そのコストは非常に大きくなる. そして, その結果が深刻であればあるほど, このコストはさらにより大きくなるであろう. 我々は, このように対象によってコストが異なり, かつ, 真のコスト関数 $c(x, y)$ は未知であるようなコスト考慮型学習の枠組みでも最も一般的な場合を扱う [14, 9, 1].

[2]通常, 報酬ではなくペナルティとして与える. 報酬の場合には, コストはマイナスになる.

$c(\mathbf{x}, h(\boldsymbol{\theta}))$ を, $\mathbf{x}$ に対して $h(\mathbf{x}, y; \boldsymbol{\theta})$ を用いてアクションを決定した場合に引き起こされるコストとする. 択一的アクションの場合 (1) には, $c(\mathbf{x}, h(\boldsymbol{\theta}))$ は

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

となる. 分散投資型のアクションの場合には, その定義は必ずしも自明ではないが, ここではもっとも簡単な場合として $c(\mathbf{x}, h(\boldsymbol{\theta}))$ が,

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

のように, それぞれのアクションからもたらされるコストが, 投資量に比例するものとする.

もしも, (2) を用いて確率的なアクション決定が可能な場合にも, (4) を用いることができるが, この場合には, $c(\mathbf{x}, h(\boldsymbol{\theta}))$ は実現するコストではなく, $\mathbf{x}$ に対する期待コストとなっていることに注意する.

2.3 コスト考慮型学習

従来, コストを考慮するような分類問題に対しては, コストの期待値を最小化することを目的とした手法が用いられてきた. 具体的には, データの分布 D に対する期待コスト

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

を最小化するように $\boldsymbol{\theta}$ を決定したいところではあるが, 実際には D は分からないので, 経験期待コスト

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

を最小化するようにパラメータ $\boldsymbol{\theta}$ を決定することになる [14, 9, 1]. 尚, 対象とコストは $X \times \mathbb{R}^Y$ で定義された確率分布 D から互いに独立に発生すると考え, D からサンプリングされた, N 個のデータの集合 E が訓練データとして与えられるとする. ここで, E の i 番目の訓練データ $\mathbf{e}^{(i)} = (\mathbf{x}^{(i)}, \{c^{(i)}(\mathbf{x}^{(i)}, y)\}_{y \in Y})$ とする. $\mathbf{x}^{(i)} \in X$ は訓練データの i 番目の対象とし, それぞれのアクション $y \in Y$ をとった場合のコスト $c^{(i)}(\mathbf{x}^{(i)}, y)$ が与えられているとする.

2.4 コスト考慮型学習の問題点

しかしながら, リスク管理の立場から考えると, 単に経験期待コストを最小化するというアプローチでは十分でない場合がある. 訓練後, M 個のデータに対してアクションをとるものとする. M が大きいときには, これらに対

するコストの和は $M \cdot C^D(\boldsymbol{\theta})$ に近づくため, $C^E(\boldsymbol{\theta})$ を学習の目的関数とすることに問題はないように思える. しかしながら, M が比較的少ないため上記の近似が成り立たず, また, 大きなコストの発生が致命的になるような状況を考えてみる. 例えば, 資金をどこに投資するべきかを決定するような問題の場合, 大きな失敗が何度か連続して起きるということは, 破産のリスクに直結する重大な問題である. 起こる確率は小さいが, 許容できないほど大きなコストが発生してしまうような可能性がある場合, ユーザーはそのリスクをなるべく回避することを望むであろう. また, たとえば, 同じだけのコスト期待値が望める 2 つの決定関数 h_1 と h_2 があったとする. h_1 によってもたらされるコストの確率分布は期待値の周りに高いピークをもつが, h_2 によってもたらされるコストの確率分布は高コストの領域に, なだらかで裾野が厚い形をもっているとする. この場合, 期待コストは等しくとも, 望ましいのは, 高いコストが発生してしまう可能性がより小さい h_1 のほうであろう.

このような場合においては, 経験期待コスト最小化では目的を正しく反映しているとはいえず, コストの分布を考慮して, リスクをより積極的に回避するような学習手法が望まれる.

3 期待ショートフォール最小化によるリスク回避型学習

前節での問題意識をもとに, 期待コストにかわる, リスク軽減を積極的に指向した目的関数を用いた学習手法を提案する.

3.1 期待ショートフォール

我々の目的は, 大きなコストが発生してしまうリスクをなるべく小さく抑えることである. 金融工学の分野では, この目的を達成するために様々なリスク指標が研究されているが, その中でももっともポピュラーであるのが, バリュー・アット・リスク (Value-at-Risk:VaR) [17] と呼ばれる量であろう. バリュー・アット・リスクとは, $0 \le \beta \le 1$ を定数として, 確率 $100\beta\%$以下の確率で生じる大きなコストの最小値（＝トップ $100\beta\%$コストの最小値）である. 我々の場合では, 意思決定関数 h の, データを発生する分布 D に対するバリュー・アット・リスク $\alpha_\beta^D(\boldsymbol{\theta})$ は,

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

のように定義される. ここで $I(\cdot)$ は引数となる条件が真のときに 1, そうでないとき 0 をとるような関数とする.

しかしながら, バリュー・アット・リスクは金融工学の分野で標準的に用いられているものの, いくつかの問題点があることが指摘されている [16, 2]. 第 1 の問題は, 2.4 節の議論と同様に, バリュー・アット・リスク以上の大きなコストが発生する場合, それがどのくらい大きくなるかはまったく考慮されていない点である. 我々の興味は, むしろ, バリュー・アット・リスク以上のコストの大きさそのものにあり, これをいかにして抑えるかというのが問題である. さらに, バリュー・アット・リスクは多くの場合凸でないことが理論的, 実験的に示されており, これは, 実際に学習を行ううえで極めて不都合である. コスト分布が正規分布などの楕円分布に従う場合には, バリュー・アット・リスクは期待コストと, コストの標準偏差の線形結合で表されるため, 上記の 2 つの問題は解決されるが, この仮定は通常成り立たない.

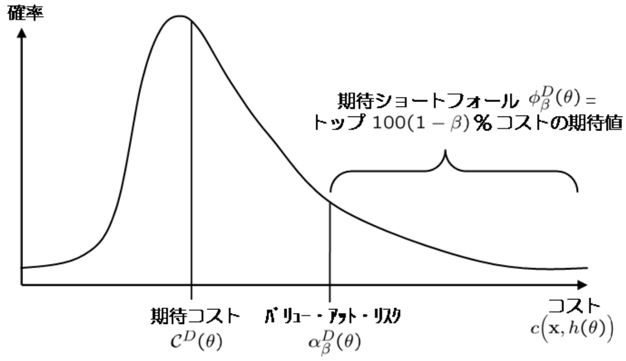

図 1: 期待コスト, バリュー・アット・リスク, 期待ショートフォール（条件付きバリュー・アット・リスク）の定義.

そこで, 我々は目的関数として, 金融工学における新しいリスク指標として近年注目を浴びている期待ショートフォール [16, 2]（あるいは条件付きバリュー・アット・リスク）を採用する. 期待ショートフォールとは, バリュー・アット・リスク以上のコストの期待値, 言い換えれば, トップ $100\beta\%$コストの期待値であり, これはまさに我々の目的と合致するリスク指標である. また, 期待ショートフォールには凸性などの望ましい性質があり [12], 学習の観点からも非常に好ましい.

さて, 期待ショートフォールを我々の枠組みにおいて定義すると, データを発生する分布 D に対する意思決定関数 $h(\mathbf{x}, y; \boldsymbol{\theta})$ の期待ショートフォール $\phi_\beta^D(\boldsymbol{\theta})$ は以下のように定義される.

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

ここで $\alpha_\beta^D(\boldsymbol{\theta})$ は前述のバリュー・アット・リスクであり, 期待ショートフォールの定義には, バリュー・アット・リスクが含まれていることに注意する.

尚, 期待ショートフォールは $\alpha_\beta^D(\boldsymbol{\theta})$ を超えるコストの期待値であることに注意すると, (7) は次のように分解することができる.

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

ここで $[x]^+ = \max\{x, 0\}$ とする.

3.2 リスク回避型学習アルゴリズム

3.2.1 経験期待ショートフォール最小化アルゴリズム

パラメータ $\boldsymbol{\theta}$ を具体的に推定するアルゴリズムを導く.

我々が最小化したい目的関数は (8) であるが, 実際には D は未知であるため, E に対して定義された経験期待ショートフォール

$$\phi_\beta^E(\boldsymbol{\theta}) = \alpha_\beta^E(\boldsymbol{\theta}) + \frac{1}{(1-\beta)N}\sum_{i=1}^N \left[c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) - \alpha_\beta^E(\boldsymbol{\theta})\right]^+, \tag{9}$$

を用いることにする. ここで $\alpha_\beta^E(\boldsymbol{\theta})$ は E に対するバリュー・アット・リスク

$$\alpha_\beta^E(\boldsymbol{\theta}) = \min\Big\{\alpha \in \mathbb{R} \mid \frac{1}{N}\sum_{i=1}^N I\big(c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) \geq \alpha\big) \leq 1-\beta\Big\} \tag{10}$$

である.

さて, (9) において, 仮に $\alpha_\beta^E(\boldsymbol{\theta})$ が既知の定数 $\tilde{\alpha}$ であるとすると, 2 項目の

$$\tilde{C}_{\tilde{\alpha}}^E(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^N \left[c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) - \tilde{\alpha}\right]^+ \tag{11}$$

だけを最小化すればよいことになる. 尚, $[x]^+$ は x について非減少な凸関数であるので, $c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ が $\boldsymbol{\theta}$ について凸であれば, (11) も同様に凸になる. ひとまずのところは, (11) を最小化するアルゴリズムがあるものと仮定する.

今度は $\boldsymbol{\theta}$ が与えられたとしよう. $\boldsymbol{\theta}$ に対するバリュー・アット・リスク (10) は訓練データについて定義されているから,

$$\alpha_\beta^E(\boldsymbol{\theta}) = \min_{k=1,\ldots,N}\Big\{c(\mathbf{x}^{(k)}, h(\boldsymbol{\theta})) \mid \frac{1}{N}\sum_{i=1}^N I\big(c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta})) \geq c(\mathbf{x}^{(k)}, h(\boldsymbol{\theta}))\big) \leq 1-\beta\Big\}$$

のように書き直すことができる. これは, $\boldsymbol{\theta}$ によって発生するコストの中で $\lfloor(1-\beta)N\rfloor$ 番目に大きいもの $c(\mathbf{x}^{(k)}, h(\boldsymbol{\theta}))$ に等しいため, 順序統計量を求めるアルゴリズムによって $O(N)$ 時間で発見することができる [4].

以上の観察より我々は, 図 2 に示す, リスク回避型学習アルゴリズム MetaRisk を提案する. このアルゴリズムは, 既存のコスト考慮型学習アルゴリズムを利用し, モデルのパラメータとバリュー・アット・リスクを交互に求めることによって経験期待ショートフォールを最小化するようなメタアルゴリズムになっている.

このアルゴリズムの最適性と収束は, 期待ショートフォールの上界の凸性を示す次の定理 [12] によって直接的に保証される.

Algorithm: MetaRisk(E, β)

- [**Step:1**] $\tilde{\alpha} := 0$ にセットする
- [**Step:2**] 現在の $\tilde{\alpha}$ において, $\tilde{C}_{\tilde{\alpha}}^E(\boldsymbol{\theta})$ を最小にする $\boldsymbol{\theta}'$ を求め, $\boldsymbol{\theta} := \boldsymbol{\theta}'$ とする
- [**Step:3**] 現在の $\boldsymbol{\theta}$ におけるバリュー・アット・リスク $\alpha_\beta^E(\boldsymbol{\theta})$ を求め, $\tilde{\alpha} := \alpha_\beta^E(\boldsymbol{\theta})$ とする
- [**Step:4**] [Step:2]〜[Step:3] を $F_\beta^E(\boldsymbol{\theta}, \tilde{\alpha})$ が収束するまで繰り返す

図 2: リスク回避型分類学習アルゴリズム MetaRisk

Theorem 1 (Rockafellar and Uryasev [12]).

$$F_\beta^E(\boldsymbol{\theta}, \alpha) = \alpha + \frac{1}{(1-\beta)N}\sum_{i=1}^N \left[c(\boldsymbol{x}^{(i)}, h(\boldsymbol{\theta})) - \alpha\right]^+, \tag{12}$$

とすると,

$$\min_{\boldsymbol{\theta}} \phi_\beta^E(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta},\alpha} F_\beta^E(\boldsymbol{\theta}, \alpha). \tag{13}$$

が成立する. また, $F_\beta^E(\boldsymbol{\theta}, \alpha)$ は α について凸であり, さらに, (6) が $\boldsymbol{\theta}$ について凸であれば $F_\beta^E(\boldsymbol{\theta}, \alpha)$ は $\boldsymbol{\theta}$ と α について凸である. さらに,

$$\alpha_\beta^E(\boldsymbol{\theta}) = \min\{\alpha \in \operatorname*{argmin}_\alpha F_\beta^E(\boldsymbol{\theta}, \alpha)\} \tag{14}$$

が成立する. □

(13) により, (12) の最小化は期待ショートフォール最小化に等しいことがわかる. また, (12) の凸性から, $\boldsymbol{\theta}$ と α についてグリーディなアルゴリズムによって最適化が可能であること, さらに (14) より, $\alpha_\beta^E(\boldsymbol{\theta})$ は $F_\beta^E(\boldsymbol{\theta}, \alpha)$ を最小化する α であるため, MetaRisk は $F_\beta^E(\boldsymbol{\theta}, \alpha)$ を $\boldsymbol{\theta}$ と α のそれぞれについて交互に最適化を行うようなアルゴリズムになっていることがわかる.

3.3 コスト考慮型からリスク回避型への変換

既存の事例依存型コスト考慮型学習アルゴリズムを利用して, (11) を最小化する方法を述べる.

3.3.1 択一型の決定関数

アクションが (1) によって択一的に決定される場合には, 実現されるコストは $[c^{(i)}(\mathbf{x}^{(i)}, y) - \tilde{\alpha}]^+ + \tilde{\alpha}$ の形に限定される. (11) は $\tilde{\alpha}$ を超えるコストの期待値になっていることに注意すると,

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

をもともとのコストに置き換えることにより, (6) は

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

となり, これは (6) と同じ形になるため, 既存の事例依存型コスト考慮型学習アルゴリズム [9, 15, 1] に, (15) のようにコストを変更した訓練例を与えることによって最小化することができる.

3.3.2 分散投資型の決定関数

次に, (2) の制約のもとで, 分散投資型あるいは確率的なアクション選択を行う場合を考える. このとき (11) は次のように書き直される.

$$\tilde{C}^E_{\tilde{\alpha}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \big[\sum_y h(\mathbf{x}^{(i)}, y; \boldsymbol{\theta}) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha} \big]^+ \quad (17)$$

択一的な場合と異なり, $c^{(i)}(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ は $c^{(i)}(\mathbf{x}^{(i)}, y)$ の線形和として表されるため, (15) のような単純な重み付けではコスト考慮型学習に帰着できない.

$h(\mathbf{x}, y; \boldsymbol{\theta})$ の自然な選択としては, たとえば多クラスのロジスティック回帰などが考えられるが, 残念ながら, $c(\mathbf{x}^{(i)}, h(\boldsymbol{\theta}))$ がパラメータについて凸にならず, 多峰性を示すため都合が悪い. そこで, 我々は (17) が $\boldsymbol{\theta}$ について線形になるようなクラスの $h(\mathbf{x}, y; \boldsymbol{\theta})$ を考え, これを勾配ブースティング [8, 13] によって推定するというアプローチをとることにする.

勾配ブースティングでは, $h(\mathbf{x}, y; \boldsymbol{\theta})$ は T 個の弱仮説 $f_1, \ldots, f_T$ の線形和として表現される.

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

ここで, $\boldsymbol{\theta}_t = (w_1, \ldots, w_t)$ はパラメータとする. また, $h(\mathbf{x}, y; \boldsymbol{\theta})$ が確率的な制約 (2) を満たすため,

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

となる.

各ブースティングラウンド t において, h_{t-1} までがすでに求まっているとすると, 新しい h_t は, h_{t-1} に確率的な制約を保存したまま弱仮説 f_t を加える形で, 次のように再帰的につくられる.

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

ここで, $0 < \gamma_t \leq 1$ はラウンド t での更新パラメータとする. そして, 最終的なパラメータは,

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

のように求めることができる.

ラウンド t における弱仮説 f_t を求めるために, γ_t が十分に小さいとして (17) の h_{t-1} の周りでのテイラー展開

$$\begin{aligned} \tilde{C}^E_{\tilde{\alpha}} = &\sum_{i=1}^{N} \big[\sum_y h_{t-1}(\mathbf{x}^{(i)}, y; \boldsymbol{\theta}_{t-1}) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha} \big]^+ \\ &+ \gamma_t \sum_{i=1}^{N} \sum_y \Bigg(\frac{\partial \big[\sum_y h_{t-1}(\mathbf{x}^{(i)}, y; \boldsymbol{\theta}_{t-1}) c(\mathbf{x}^{(i)}, y) - \tilde{\alpha} \big]^+}{\partial h_{t-1}(\mathbf{x}^{(i)}, y; \boldsymbol{\theta}_{t-1})} \\ &\cdot \big(f_t(\mathbf{x}^{(i)}, y) - h_{t-1}(\mathbf{x}^{(i)}, y) \big) \Bigg) + O(\gamma_t^2) \end{aligned}$$

を行い, 2 次以上の項を無視すると, 第 2 項

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

を最小化する f_t を求めればよいことになる. これは, 択一的アクションの場合と同様に, (15) を

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

のように修正した訓練例 $\tilde{E}$ を, コスト考慮型のアルゴリズムに渡すことによって最小化できる.

尚, 更新パラメータ γ_t は, (17) を最小化するように, 線形探索や線形計画法によって求めることができる.

4 実験

提案法の効果を確かめるために, 計算機実験を行った.

データは, [9] で用いられている German Credit Data Set [11] を用いた[3]. このタスクは顧客の信用リスクを予測する問題で, 顧客の情報をもとに, その人がよい顧客か悪い顧客かを分類するというものである. 顧客情報 $\mathbf{x}$ の属性としては, 性別や職業, 使用目的や過去の履歴など, 20 の項目からなっている. この実験では, データセットに付属している, すべての属性を 24 の数値属性に変換したものを用いた. クラスは"good customer" (700 人) と"bad customer" (300 人) の 2 種類である. "good customer"を"bad customer"と誤って分類してしまうと, 本来入るはずだった利息分の利益を失い, 逆に"bad customer"を"good customer"と誤分類してしまうと, 貸付分の大部分を失ってしまう. このデータセットでは, もともと個々のデータのコストは与えられていないが, [9] に倣って, "good customer"を "bad customer"に誤分類してしまうコストを $0.1 \times \frac{\text{期間}}{12} \times$ 貸付額 (年率 10%の利息. 平均 6.27, 分散 43.51^2, 最大値 78.27.), また, "bad customer"を"good customer"に誤分類してしまうコストを $0.75 \times$ 貸付額 (75%が失われる. 平均 29.54, 分散 78.09^2, 最大値 138.18) とした[4]. アクションを択一的に選択する場合のモデル $h(\mathbf{x}, y)$ および分散投資型のアクションの場合の弱仮説 $f_t(\mathbf{x}, y)$ として用いるコスト考慮型学習器としてはコスト考慮型パーセプトロン [9] をカーネル化したものを用いた. カーネル関数としてはガウシアンカーネル ($\sigma = 50$) を用いた[5].

[3]データは UCI Machine Learning repository [10] より入手可能.

[4]択一的アクションが「(良い顧客と判断して) 貸し付ける」か「(悪い顧客と判断して) 貸し付けない」のどちらかを選択するのに対し, 分散投資型アクションの場合の解釈としては, 貸付額のうち, どれだけの割合を貸し付けるかということに対応する. このときのコストは実際, (4) の形になる.

[5]パラメータは予備実験において, 期待コストを最小化するようなパラメータを求めた.

表 1: 択一的アクションの場合の実験結果： 数値は期待ショートフォール, 括弧内は VaR, 最下段は平均コスト.

Test ES (VaR)	**Cost-Sensitive**	**Risk-Sensitive**
β =0.01	64.47 (46.59)	55.34 (40.66)
β =0.05	34.66 (17.67)	30.13 (15.28)
β =0.10	23.58 (10.16)	23.04 (9.09)
β =0.20	14.71 (3.30)	14.40 (3.66)
Mean Cost	3.31	-

表 1 には択一的なアクションの場合の実験結果を, 表 2 には分散投資型アクションの場合の実験結果を示す. すべての結果は, データを 3 分割（訓練データ 666 個, テストデータ 334 個）した交差検定による値の平均値によって計測した. Cost-Sensitive の列は, 期待コストを目的関数とするような従来のコスト考慮型パーセプトロンによる結果を, Risk-Sensitive の列は, 期待ショートフォールを目的関数とするような提案手法 ($\beta = 0.20, 0.10, 0.05, 0.01$ それぞれの場合) による結果を示した. 各行はテストデータにおける各 β での期待ショートフォールの 3 回の平均を示している. また, 括弧内の数値はそれぞれの場合でのバリュー・アット・リスクを示している. Cost-Sentitive の列の最下行の Mean Cost の行は平均コストを示している.

分散投資型のアクションは, 2 つのアクションのコストの間のコストが実現できるため, 択一的アクションの場合より全体的により小さなコストが実現できている. また, 期待通り, コスト考慮型学習の場合には対応する β における期待ショートフォールを減少できていることがわかる. β が大きくなるほどリスク回避型による期待ショートフォールの減少分は著しくなくなっているが, これは, コスト分布が左側（0 付近）に大きく偏っているため, 期待ショートフォールと, 期待コストの違いが小さくなっているためと思われる. また, $\beta = 0.20$ では, コスト考慮型のほうがリスク回避型よりも小さなバリュー・アット・リスクを実現していることがみてとれる. これはバリュー・アット・リスクが小さいことは, かならずしも小さな確率で起こる大きなコストそのものを抑えることにはならないことを示唆している.

表 2: 分散投資型の場合の実験結果： 数値は期待ショートフォール, 括弧内は VaR, 最下段は平均コスト.

Test ES (VaR)	**Cost-Sensitive**	**Risk-Sensitive**
β =0.01	64.47 (46.59)	44.29 (30.69)
β =0.05	34.66 (17.67)	26.17 (14.13)
β =0.10	23.58 (10.16)	20.25 (8.95)
β =0.20	14.71 (3.30)	13.01 (3.74)
Mean Cost	3.31	-

参考文献

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

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

[5] 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.

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

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

[8] J. H. Friedman. Greegy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 2001.

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

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

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

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

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

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

[15] 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.

[16] 山井, 吉羽. リスク指標の性質に関する理論的考察 - VaR と期待ショートフォールの比較分析. **金融研究**, (12):95–132, 2001.

[17] 野口, 藤井. **金融工学-ポートフォリオ選択と派生資産の経済分析**. ダイヤモンド社, 2000.

[18] 鈴木. 正確な学習よりも得する学習 - 誤分類コストを考慮する分類学習 - (1)(2). **情報処理**, 45(4-5), 2004.