Innovative R&D by NTT

Learning Latest Classifiers without Additional Labeled Data (IJCAI 2017)

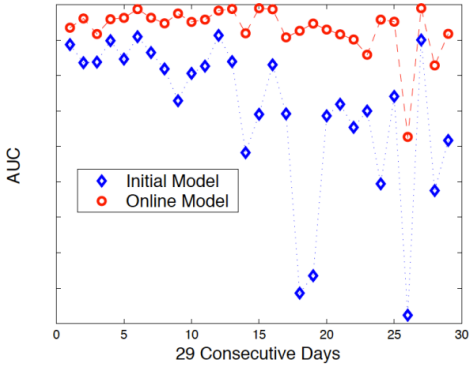
熊谷　充敏（NTTセキュアプラットフォーム研究所）
岩田　具治（NTTコミュニケーション科学基礎研究所）

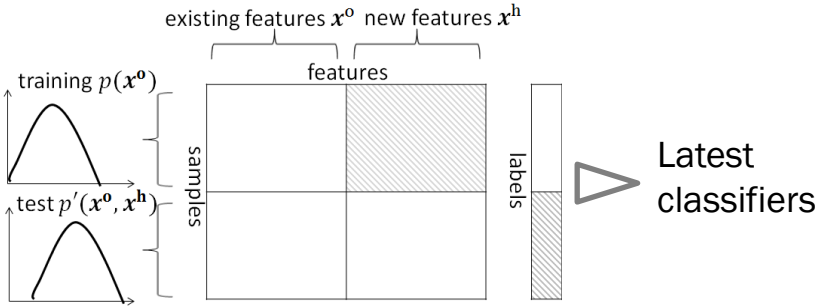
Copyright©2017 NTT corp. All Rights Reserved.

背景

- 多くのドメインで分類器の性能は経年劣化
  - セキュリティ、IoT、マーケティング、…
- ラベルありデータによる分類器の更新が有効
  ⇒しかし、ラベルありデータの収集は高コスト

- 本研究では、ラベルありデータに加え、ラベルなしデータを用いて性能劣化を抑える手法を考える

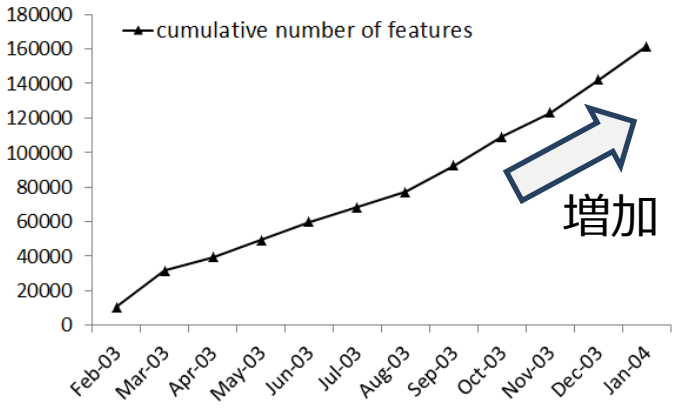




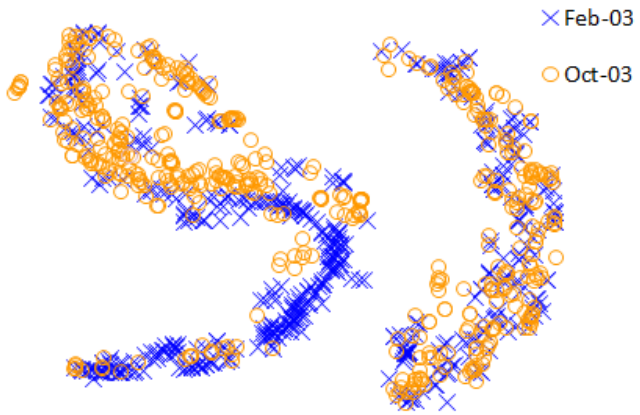
Copyright©2017 NTT corp. All Rights Reserved.

経年劣化の２つの主要因

- 分類器の経年劣化の要因は主に２つ
  - 新たな特徴(features)の出現
  - データ分布の変化



スパムメールにおける特徴の増加

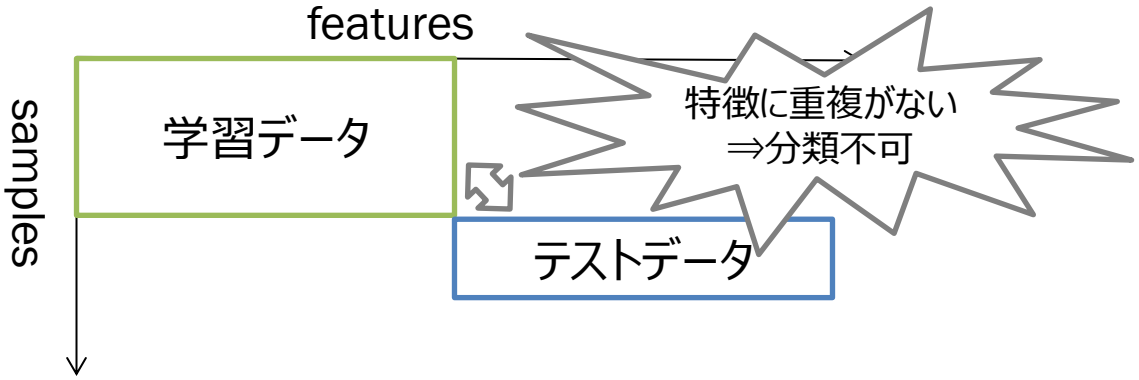


スパムのデータ分布の変化
(時期の異なるスパムをt-SNEでマッピング)

NTT
Copyright©2017 NTT corp. All Rights Reserved.

新たな特徴の出現

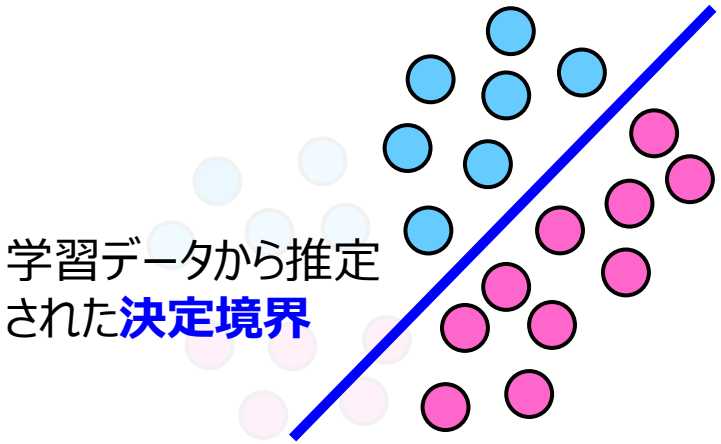
- “特徴”の数が時間とともに増えることがある
  - スパムメール分類：
    スパムメールを表す単語(特徴)が増える
  - センサーデータマイニング：
    新たなセンサー(特徴)が設置される
- 学習データに現れなかった“特徴”が増えると、分類性能が落ちる



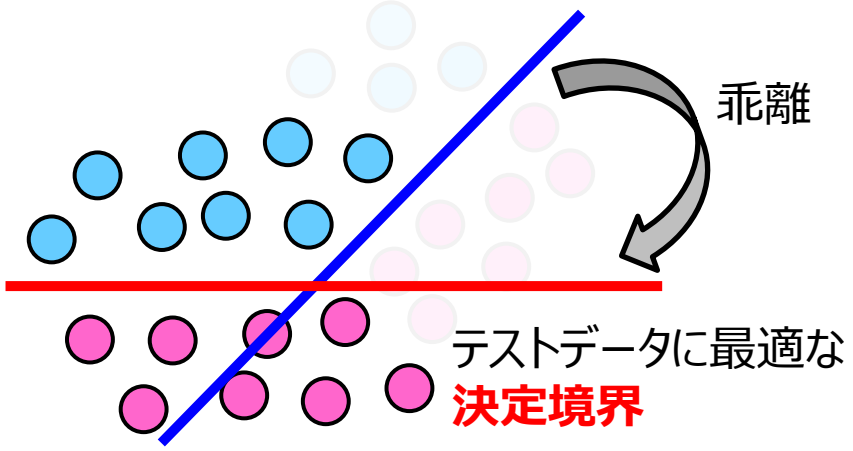
NTT
Copyright©2017 NTT corp. All Rights Reserved.

データ分布の変化

- 学習/テストでデータ分布が変化することがある
  - スパムメール,マルウェア,センサーデータ,BCI,…
- データ分布が変化すると、最適な決定境界も変わるため、分類性能が落ちる



学習フェーズ



テストフェーズ

Copyright©2017 NTT corp. All Rights Reserved.

本研究の概要

- ラベルなしデータを活用し、“新たな特徴の出現”と“データ分布の変化”という２つの問題を同時に解決する分類器学習法を提案

- それぞれの問題を解ける研究はいろいろある
  - D. Williams et al, Incomplete-data classification using logistic regression,ICML2005
  - H.Shimodaira et al, Improving predictive inference under covariate shift by weighting the log-likelihood function, JSPI2000
- 両問題を明示的に扱い、かつ同時に解く点が新しいところ

Copyright©2017 NTT corp. All Rights Reserved.

タスク（1/3）

入力：

- ある時刻までに得られたラベル**あり**データ：

$$\mathcal{D} := \{(\boldsymbol{x}_n^{\mathrm{o}}, y_n)\}_{n=1}^{N}$$

n番目の$\boldsymbol{D_o}$**次元**特徴ベクトル　　クラスラベル（0 or 1）

- それ以降の時刻に得られたラベル**なし**データ(テストデータ)：

$$\mathcal{D}' := \{(\boldsymbol{x}_m^{\mathrm{o}}, \boldsymbol{x}_m^{\mathrm{h}})\}_{m=1}^{M}$$

m番目の$\boldsymbol{D_o + D_h}$**次元**特徴ベクトル

注：ラベルなしデータはラベルありデータの後に収集されるため、
ラベルなしデータには新しい特徴$\boldsymbol{x}^{\mathrm{h}}$が含まれる

NTT
Copyright©2017 NTT corp. All Rights Reserved.

タスク（2/3）データ分布の変化に対する仮定

- 学習/テストのデータ分布の変化のモデルとして、
  共変量シフト(covariate shift)の仮定をおく

$$p(x) \neq p'(x), \qquad p(y|x) = p'(y|x)$$

x：特徴ベクトル
y：クラスラベル

p：学習時の分布
p'：テスト時の分布

- この状況は多くの実問題で現れることが報告されている
  - スパムメール分類
  - HIV検査
  - センサーデータによる行動認識

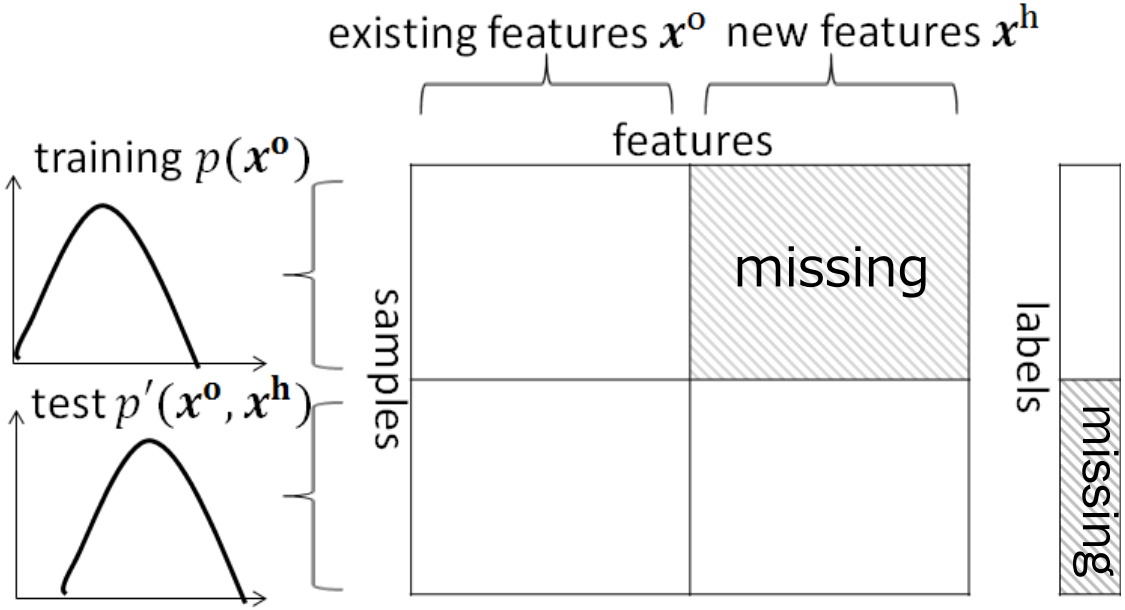
NTT
Copyright©2017 NTT corp. All Rights Reserved.

タスク（3/3）

出力：

テスト分布$p'(\boldsymbol{x}^{\mathrm{o}}, \boldsymbol{x}^{\mathrm{h}})$からのサンプルを精度よく分類できる分類器

$$h : \mathbb{R}^{D_{\mathrm{o}}+D_{\mathrm{h}}} \to \{0, 1\}$$



本タスクで取り扱うデータ

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (1/9)

- テスト分布における汎化誤差Gを最小化する分類器hを求めたい

$$G := \int \underbrace{\mathrm{loss}(y, h(\boldsymbol{x}^{\mathrm{o}}, \boldsymbol{x}^{\mathrm{h}}))}_{\text{任意のロス関数}} \underbrace{p'(\boldsymbol{x}^{\mathrm{o}}, \boldsymbol{x}^{\mathrm{h}}, y)}_{\text{テスト分布}} d\boldsymbol{x}^{\mathrm{o}} d\boldsymbol{x}^{\mathrm{h}} dy,$$

- 通常はテスト分布からのサンプルでGを近似して求める
- 今はテスト分布からラベル y のサンプルが得られない
  ⇒単純には求められない

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (2/9)

- 仮定$p(y|\boldsymbol{x}^{\mathrm{o}},\boldsymbol{x}^{\mathrm{h}}) = p'(y|\boldsymbol{x}^{\mathrm{o}},\boldsymbol{x}^{\mathrm{h}})$を使うと、Gはテスト分布のラベルyを使わずに書ける

$$G = \int \mathrm{loss}(y, h(\boldsymbol{x}^{\mathrm{o}},\boldsymbol{x}^{\mathrm{h}}))\frac{p'(\boldsymbol{x}^{\mathrm{o}},\boldsymbol{x}^{\mathrm{h}})}{p(\boldsymbol{x}^{\mathrm{o}},\boldsymbol{x}^{\mathrm{h}})}p(\boldsymbol{x}^{\mathrm{o}},\boldsymbol{x}^{\mathrm{h}},y)d\boldsymbol{x}^{\mathrm{o}}d\boldsymbol{x}^{\mathrm{h}}dy$$

学習分布

- さらに以下の仮定をおく

$$p(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}}) = p'(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}})$$

学習　　テスト

- ラベルありデータには$\boldsymbol{x}^{\mathrm{h}}$はない
  ⇒学習時の分布$\mathrm{p}(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}})$を直接知ることは不可能
- そのため、学習/テストで両分布が等しいと仮定

NTT
Copyright©2017 NTT corp. All Rights Reserved.

提案法 (3/9)

- 汎化誤差Gは、ラベルありデータDの重み付き経験損失で近似できる

$$G \approx \frac{1}{N} \sum_n \tau(\boldsymbol{x}_n^o) \int \mathrm{loss}(y_n, h(\boldsymbol{x}_n^o, \boldsymbol{x}_n^h)) p(\boldsymbol{x}_n^{\mathrm{h}} | \boldsymbol{x}_n^{\mathrm{o}}) d\boldsymbol{x}_n^h,$$

$$\tau(\boldsymbol{x}^{\mathrm{o}}) := \frac{p'(\boldsymbol{x}^{\mathrm{o}})}{p(\boldsymbol{x}^{\mathrm{o}})}$$

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (4/9)

- 汎化誤差Gは、ラベルありデータDの重み付き経験損失で近似できる

$$G \approx \frac{1}{N} \sum_{n} \tau(\boldsymbol{x}_n^{o}) \int \mathrm{loss}(y_n, h(\boldsymbol{x}_n^{o}, \boldsymbol{x}_n^{h})) p(\boldsymbol{x}_n^{\mathrm{h}} | \boldsymbol{x}_n^{\mathrm{o}}) d\boldsymbol{x}_n^{h},$$

ラベルありデータ$\boldsymbol{x}_n^{\mathrm{o}}$における
未観測の特徴x_n^{h}を予測する分布

$$\tau(\boldsymbol{x}^{\mathrm{o}}) := \frac{p'(\boldsymbol{x}^{\mathrm{o}})}{p(\boldsymbol{x}^{\mathrm{o}})}$$

ラベルありデータ$\boldsymbol{x}^{\mathrm{o}}$の分類器学習における
重要度(密度比)を表す量

- 条件付き分布$p(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}})$と密度比$\tau(\boldsymbol{x}^{\mathrm{o}})$が決まれば、分類器hを学習できる

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (5/9) 条件付き分布

- $p(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}})$のモデルとしてガウス分布を仮定

$$p(\boldsymbol{x}_n^{\mathrm{h}}|\boldsymbol{x}_n^{\mathrm{o}}) = \mathcal{N}\left(\boldsymbol{x}_n^{\mathrm{h}} \,\middle|\, \boldsymbol{A}\boldsymbol{x}_n^{\mathrm{o}} + \boldsymbol{a}, \boldsymbol{\Lambda}^{-1}\right),$$
$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C},\ \boldsymbol{B} \in \mathbb{R}^{D_{\mathrm{h}} \times K},\ \boldsymbol{C} \in \mathbb{R}^{K \times D_{\mathrm{o}}},$$
$$\Lambda = \mathrm{diag}(\boldsymbol{\lambda})^2 + \epsilon \boldsymbol{I}_{\mathrm{D_h}}$$

- パラメータ$\mathbf{B}, \mathbf{C}, \mathbf{a}, \boldsymbol{\lambda}$は、仮定$p\left(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}}\right) = p'\left(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}}\right)$により、ラベルなしデータだけで推定可能

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (6/9) 条件付き分布

- MAP推定でパラメータを推定

尤度　ガウス分布　ガウス分布　ガンマ分布

$$
\begin{aligned}
F &= \log p(\mathcal{D}'|\boldsymbol{B},\boldsymbol{C},\boldsymbol{a},\boldsymbol{\lambda}) + \log p(\boldsymbol{B}) + \log p(\boldsymbol{C}) + \log p(\boldsymbol{\lambda}^2) \\
&= -\frac{1}{2}\sum_m (\boldsymbol{x}_m^{\mathrm{h}} - \boldsymbol{B}\boldsymbol{C}\boldsymbol{x}_m^{\mathrm{o}} - \boldsymbol{a})^\top \boldsymbol{\Lambda} (\boldsymbol{x}_m^{\mathrm{h}} - \boldsymbol{B}\boldsymbol{C}\boldsymbol{x}_m^{\mathrm{o}} - \boldsymbol{a}) \\
&\quad - \frac{r}{2}\mathrm{Tr}(\boldsymbol{B}^\top \boldsymbol{B}) - \frac{r}{2}\mathrm{Tr}(\boldsymbol{C}^\top \boldsymbol{C}) + \frac{1}{2}\log\det(\boldsymbol{\Lambda}) \\
&\quad + \sum_k 2(a-1)\log\lambda_k - b\lambda_k^2 + const,
\end{aligned}
$$

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (7/9) 密度比$\tau(\boldsymbol{x}^{\mathrm{o}}) = \frac{p'(\boldsymbol{x}^{\mathrm{o}})}{p(\boldsymbol{x}^{\mathrm{o}})}$

- ナイーブな方法
  - $p'(\boldsymbol{x}^{\mathrm{o}})$と$p(\boldsymbol{x}^{\mathrm{o}})$をそれぞれ独立に推定し、比をとる
  - 比を取ることで推定誤差増大の恐れあり

- **密度比推定法**
  - 密度比$\frac{p'(\boldsymbol{x}^{\mathrm{o}})}{p(\boldsymbol{x}^{\mathrm{o}})}$を“直接”推定する方法
  - ナイーブな方法より高性能
  - 様々な方法が提案

Copyright©2017 NTT corp. All Rights Reserved.

提案法 (8/9) 密度比$\tau(\boldsymbol{x}^{\mathrm{o}}) = \frac{p'(\boldsymbol{x}^{\mathrm{o}})}{p(\boldsymbol{x}^{\mathrm{o}})}$

Innovative R&D by NTT
NTT Confidential

- **Unconstrained LSIF**[1]
  - 密度比τを線形モデルで表す

$$\hat{\tau}(\boldsymbol{x}^{\mathrm{o}}) = \boldsymbol{\alpha}^{\top}\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{o}}) = \sum_{m=1}^{M} \alpha_m \exp\left(-\frac{\|\boldsymbol{x}^{\mathrm{o}} - \boldsymbol{x}_m^{o}\|^2}{2\gamma^2}\right),$$

  - パラメータ$\boldsymbol{\alpha}$は目的関数Jの解析解として得られる

$$J(\boldsymbol{\alpha}) = \int (\tau(\boldsymbol{x}^{\mathrm{o}}) - \hat{\tau}(\boldsymbol{x}^{\mathrm{o}}))^2 p(\boldsymbol{x}^{\mathrm{o}}) d\boldsymbol{x}^{\mathrm{o}} + \underline{\rho\|\boldsymbol{\alpha}\|^2},$$

正則化項

$$\boldsymbol{\alpha} = \left(\frac{1}{N}\sum_n \psi(\boldsymbol{x}_n^{\mathrm{o}})\psi(\boldsymbol{x}_n^{\mathrm{o}})^{\top} + \rho\boldsymbol{I}\right)^{-1}\left(\frac{1}{M}\sum_m \psi(\boldsymbol{x}_m^{\mathrm{o}})\right).$$

1 T.Kanamori et al, Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection, NIPS2009

NTT
Copyright©2017 NTT corp. All Rights Reserved.

提案法 (9/9)

- 推定した条件付き分布$p(\boldsymbol{x}^{\mathrm{h}}|\boldsymbol{x}^{\mathrm{o}})$と密度比$\tau(\boldsymbol{x}^{\mathrm{o}})$を経験損失Gに代入
- 識別モデルとしてロジスティック回帰を採用

$$\mathrm{p}(y|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{1+\exp(-\boldsymbol{x}\cdot\boldsymbol{w})},\ \boldsymbol{w} = (\boldsymbol{w}_{\mathrm{o}}, \underline{\boldsymbol{w}_{\mathrm{h}}}) \in \mathbb{R}^{\boldsymbol{D_o}+\boldsymbol{D_H}}$$

新特徴$\boldsymbol{x}^{\mathrm{h}}$に対応する
分類器のパラメータ

- 分類器のパラメータ$\boldsymbol{w} = (\boldsymbol{w}_{\mathrm{o}}, \boldsymbol{w}_{\mathrm{h}})$は勾配ベースの最適化法で学習可能

NTT
Copyright©2017 NTT corp. All Rights Reserved.

実験

Innovative R&D by NTT
NTT Confidential

データセット

- Spam: スパム/ハムメールの公開データセット[1]
- URL-reputation：悪性/良性サイトの公開データセット[2]
- Malicious-proxy：悪性/良性の通信データ

比較手法

- LR：ラベルありデータのみで学習
- ILR：$p(x^{\mathrm{h}}|x^{\mathrm{o}})$で新特徴を補完して学習　特徴増加のみ対応
- IWLR：密度比推定uLSIFによる分類器学習　分布変化のみ対応
- MER：ロジスティック回帰の半教師あり学習版　特徴増加のみ対応

実験設定

- ある時刻までのデータをラベルありデータ、それ以降のデータをラベルなしデータとテストデータに設定
- 分割時刻を変えながら10個のデータセットを生成し，AUCの平均で評価

1 http://www.comp.dit.ie/sjdelany/Dataset.htm
2 https://archive.ics.uci.edu/ml/datasets/URL+Reputation

NTT

実験：
ラベルあり(N個)とラベルなし(M個)の比率を変えた場合の**特徴数**

NTT Confidential

Table 2: Average number of existing features D_o and new features D_h with N labeled samples and M unlabeled samples for each data set

Dataset	N/M	D_o	D_h
spam	0.060	1895	14521
	0.125	3221	13881
	0.250	5168	13031
	0.500	8288	12119
	1.000	14915	10983
URL-reputation	0.060	1209	8900
	0.125	2020	8538
	0.250	3435	8040
	0.500	5762	7437
	1.000	9677	6740
malicious-proxy	0.060	3008	23195
	0.125	4883	22127
	0.250	8290	20299
	0.500	13247	18419
	1.000	20670	15916

Copyright©2017 NTT corp. All Rights Reserved.

結果:
ラベルあり(N個)とラベルなし(M個)の比率を変えた場合のAUC

Innovative R&D by NTT
NTT Confidential

Proopsed

N/M	Proposed	LR	ILR	IWLR	MER
0.060	0.6918 (0.0463)	0.6969 (0.0434)	**0.7125** (0.0403)	0.7095 (0.0398)	0.6926 (0.0416)
0.125	**0.7646** (0.0509)	0.7637 (0.0433)	0.7632 (0.0430)	0.7625 (0.0437)	0.7602 (0.0429)
0.250	**0.8188** (0.0346)	0.8045 (0.0424)	0.8047 (0.0422)	0.8060 (0.0412)	0.8125 (0.0409)
0.500	**0.8732** (0.0229)	0.8335 (0.0448)	0.8390 (0.0422)	0.8412 (0.0413)	0.8443 (0.0405)
1.000	**0.9154** (0.0168)	0.8747 (0.0331)	0.8677 (0.0345)	0.8708 (0.0340)	0.8721 (0.0324)
Average	**0.8128** (0.0194)	0.7947 (0.0198)	0.7974 (0.0190)	0.7980 (0.0191)	0.7964 (0.0194)

(a) spam

N/M	Proposed	LR	ILR	IWLR	MER
0.060	**0.9629** (0.0062)	0.9586 (0.0081)	0.9596 (0.0072)	0.9594 (0.0078)	0.9560 (0.0086)
0.125	**0.9770** (0.0034)	0.9749 (0.0040)	0.9765 (0.0035)	0.9752 (0.0039)	0.9749 (0.0040)
0.250	**0.9835** (0.0081)	0.9826 (0.0077)	0.9826 (0.0076)	0.9826 (0.0077)	0.9828 (0.0025)
0.500	0.9837 (0.0033)	0.9834 (0.0031)	0.9834 (0.0031)	0.9835 (0.0031)	**0.9839** (0.0032)
1.000	**0.9862** (0.0033)	0.9855 (0.0034)	0.9856 (0.0034)	0.9856 (0.0034)	0.9858 (0.0035)
Average	**0.9786** (0.0021)	0.9770 (0.0025)	0.9775 (0.0023)	0.9773 (0.0024)	0.9767 (0.0026)

(b) URL-reputation

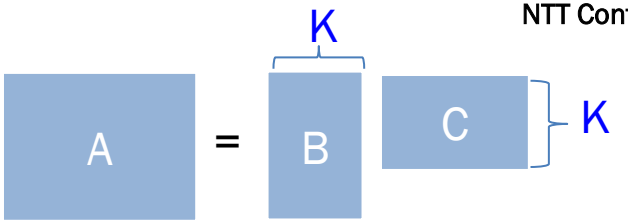
N/M	Proposed	LR	ILR	IWLR	MER
0.060	**0.5206** (0.0099)	0.5041 (0.0088)	0.5127 (0.0120)	0.5023 (0.0081)	0.5110 (0.0067)
0.125	**0.5703** (0.0138)	0.5500 (0.0108)	0.5671 (0.0134)	0.5511 (0.0111)	0.5541 (0.0149)
0.250	0.6098 (0.0184)	0.6067 (0.0191)	**0.6133** (0.0177)	0.6060 (0.0194)	0.6072 (0.0196)
0.500	0.6326 (0.0050)	0.6327 (0.0057)	0.6327 (0.0048)	0.6319 (0.0053)	**0.6335** (0.0059)
1.000	**0.7041** (0.0017)	0.6802 (0.0218)	0.6986 (0.0191)	0.6830 (0.0200)	0.6787 (0.0220)
Average	**0.6075** (0.0106)	0.5947 (0.0108)	0.6049 (0.0108)	0.5948 (0.0108)	0.5969 (0.0107)

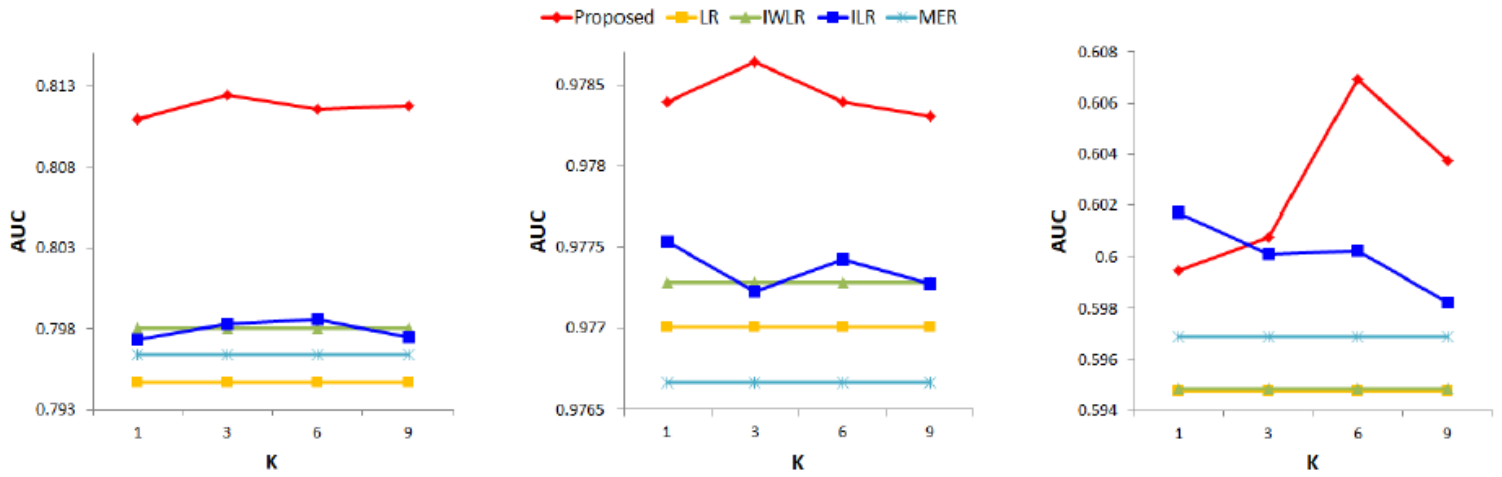
(c) malicious-proxy

多くのケース(11/15)で提案法が最高性能を達成

結果：行列分解のパラメータK依存性

$$p(\boldsymbol{x}_n^{\mathrm{h}}|\boldsymbol{x}_n^{\mathrm{o}}) = \mathcal{N}\left(\boldsymbol{x}_n^{\mathrm{h}} \,\middle|\, \boldsymbol{A}\boldsymbol{x}_n^{\mathrm{o}} + \boldsymbol{a}, \boldsymbol{\Lambda}^{-1}\right),$$
$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C},\ \boldsymbol{B} \in \mathbb{R}^{D_{\mathrm{h}} \times K},\ \boldsymbol{C} \in \mathbb{R}^{K \times D_{\mathrm{o}}},$$





(a) spam (b) URL-reputation (c) malicious-proxy

Figure 6: Average of AUC over the N/M when varying value of imputation parameter K

- Kの増加⇒AUC向上ではない
- 分類性能の観点からは小さいKでも十分であることが示唆される

NTT
Copyright©2017 NTT corp. All Rights Reserved.

まとめ

- 分類器の性能劣化の要因として、“新たな特徴量の増加”と“データ分布の変化”を仮定
- ラベルあり/なしデータを使って両問題を解決する分類器学習法を提案
- 実データを使った実験で有効性を確認

Copyright©2017 NTT corp. All Rights Reserved.