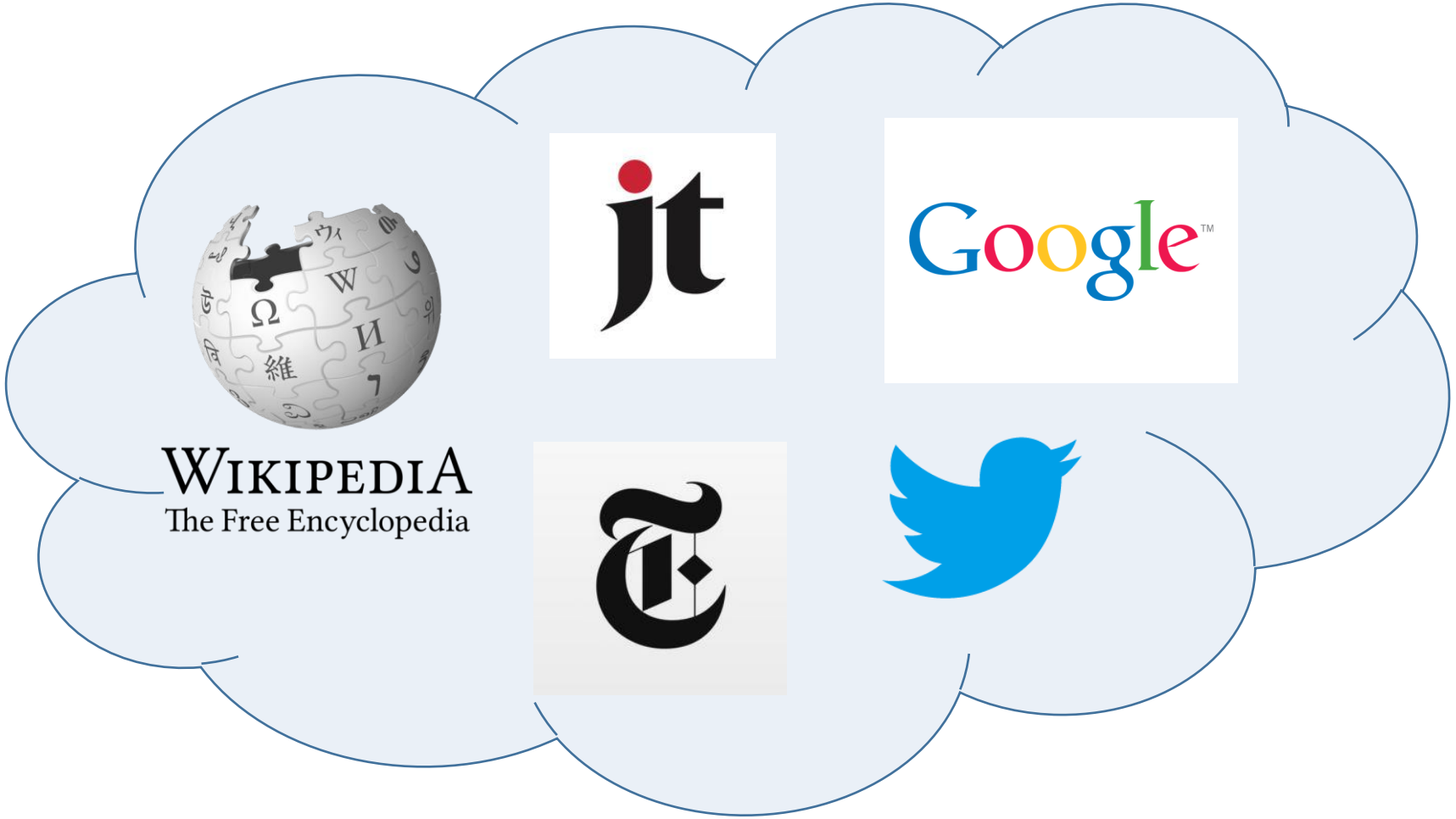
2019.3.5 DEIM @ 長崎

動的クランピングを用いた多クラス分類器

○ 澄川靖信
首都大学東京

宮崎辰郎
東京理科大学

背景｜分類の重要性



自動的にラベル付けする分類器の重要性は増加

背景｜データセット作成のコスト減

データセット作成の環境は改善している。
しかし、多ラベル分類では未だ困難がある。

背景｜ラベル欠損の存在

Category:Natural disasters

From Wikipedia, the free encyclopedia

Subcategories

This category has the following 27 subcategories, out of

A

- Avalanches (5 C, 8 P)

E

- Earthquakes (19 C, 21 P)

1974 Lesser Antilles earthquake

From Wikipedia, the free encyclopedia

The **1974 Lesser Antilles earthquake** occurred at 05:50:58 local time on October 8 with a m
were injured in what the United States' National Geophysical Data Center called a moderately c

Contents [hide]

1 Tectonic setting
2 See also
3 References
4 External links

Categories: 1974 earthquakes | 1974 in the Caribbean | Earthquakes in Antigua and Barbuda

2009 Schalfkogel avalanche

From Wikipedia, the free encyclopedia

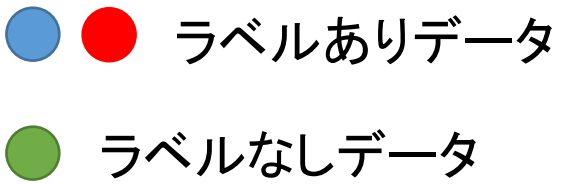
The **2009 Schalfkogel avalanche** was an avalanche which occurred in Sölden, Austria, on 2 May 2009. Six people were killed, five Czechs and o
3,500-metre (11,500 ft) Schalfkogel mountain range. The corpses were discovered to have been frozen upon recovery.[1] It was the deadliest ava
2000.[2] Although avalanches are a regular occurrence in the region, they mainly kill individuals as opposed to entire groups.[3]

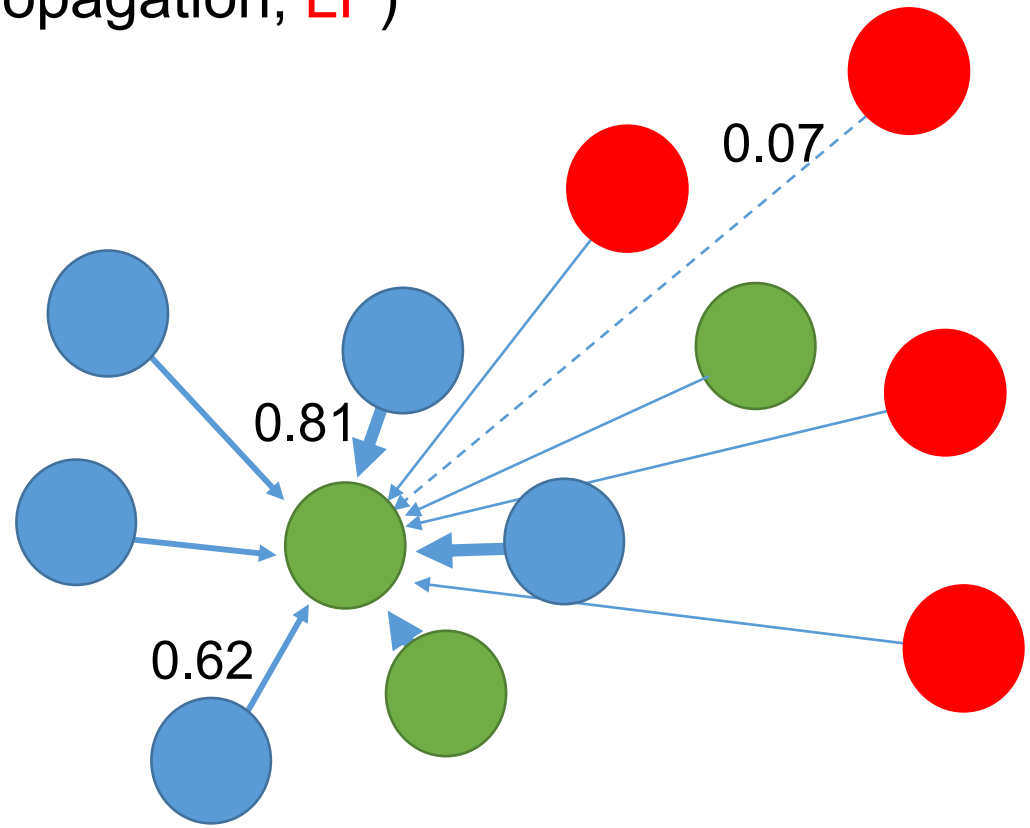
Contents [hide]

Categories: 2009 natural disasters | 2009 in Austria | 2000s avalanches | Natural disasters in Austria | May 2009 events | Avalanches in Austria

目標1｜更なるデータセット作成コスト減

- 半教師あり学習
- ラベル伝播（Lapel Propagation, LP)





値が収束するまで繰り返し伝播する。

目標2｜教師データにラベル欠損が含んでも良い結果を得る

- 多ラベル分類におけるデータセットの問題

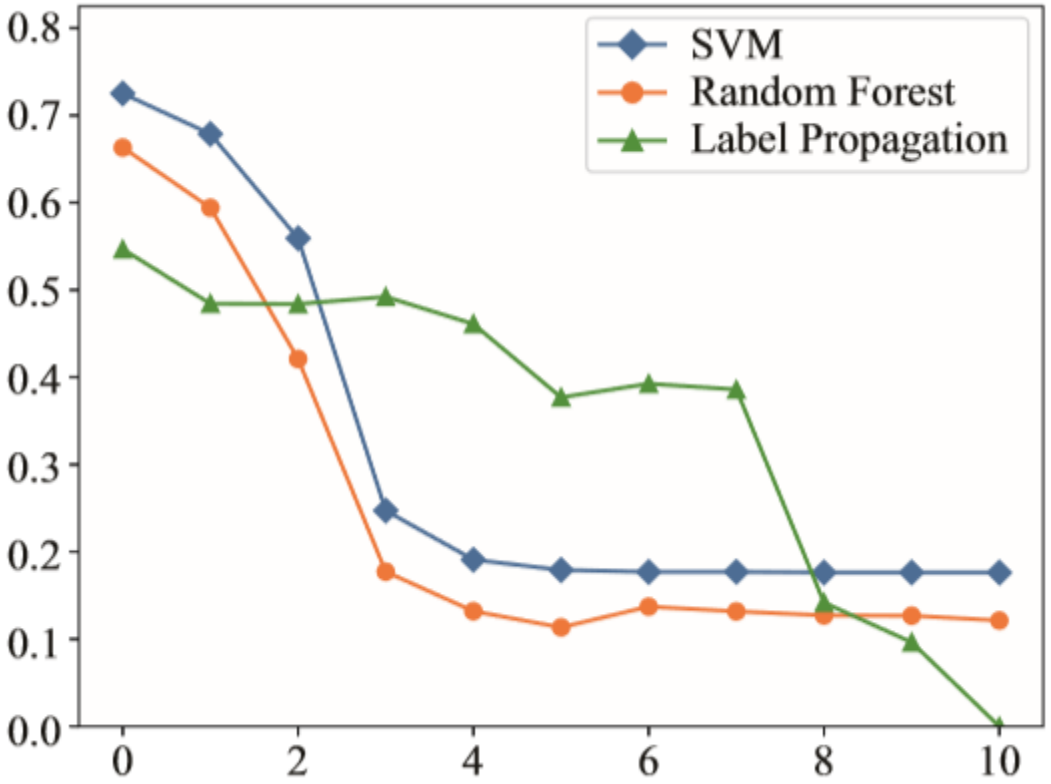
- 負例の存在：誤ったラベルの影響
  - 先行研究：Linear Neighborhood Propagation, LNP

- ラベル欠損：正しいが付与されていないラベルの影響
  - 本研究で解決する問題。
  - 本研究では負例は存在しないものと仮定する。

ラベル欠損は精度を低減する

SIAM2007テキストマイニングコンペティションデータセットの例

マイクロ
平均F値

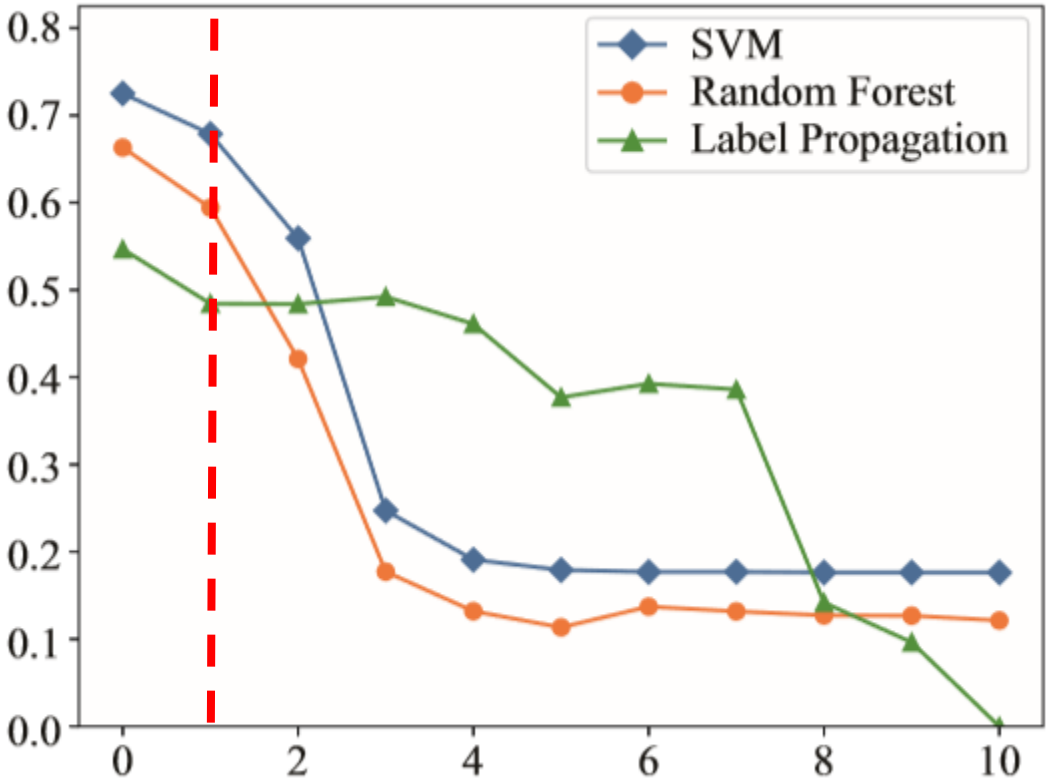


ラベル欠損数

ラベル欠損は精度を低減する

SIAM2007テキストマイニングコンペティションデータセットの例

マイクロ
平均F値

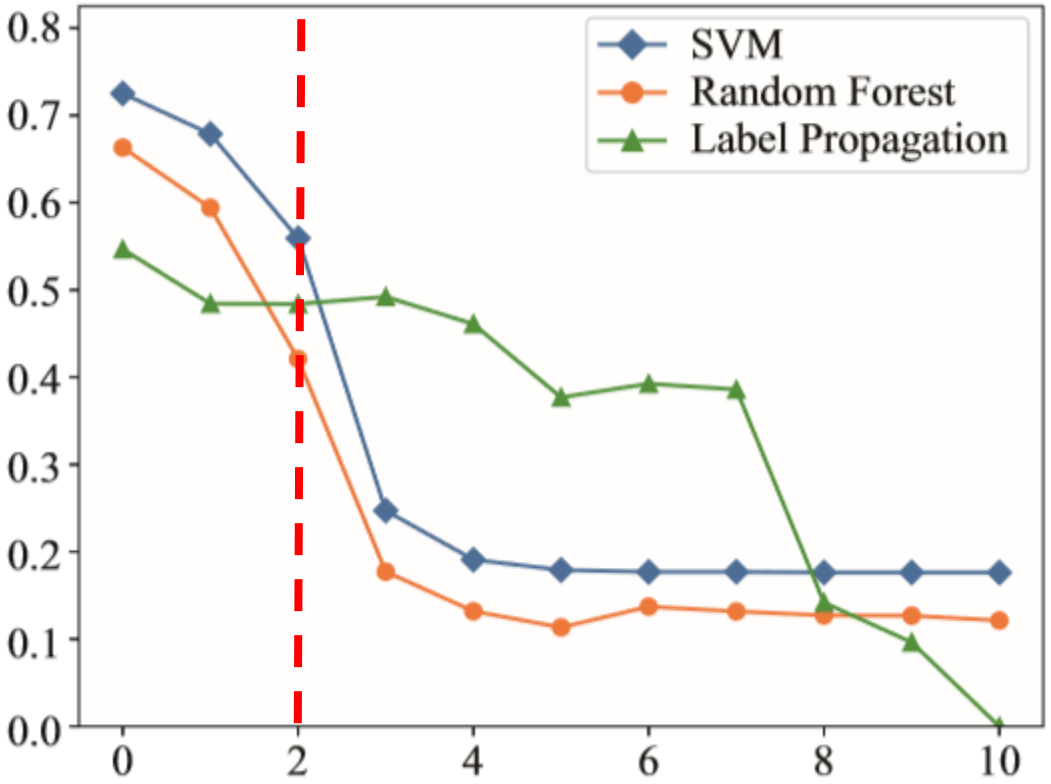


ラベル欠損数

ラベル欠損は精度を低減する

SIAM2007テキストマイニングコンペティションデータセットの例

マイクロ
平均F値



ラベル欠損数

提案手法

- Label Propagation using Amendable Clamping (LPAC).
  - 目標：教師データ中のラベル欠損による悪影響の低減。

- 実現方法
  1. 局所的伝播：伝播方法の拡張

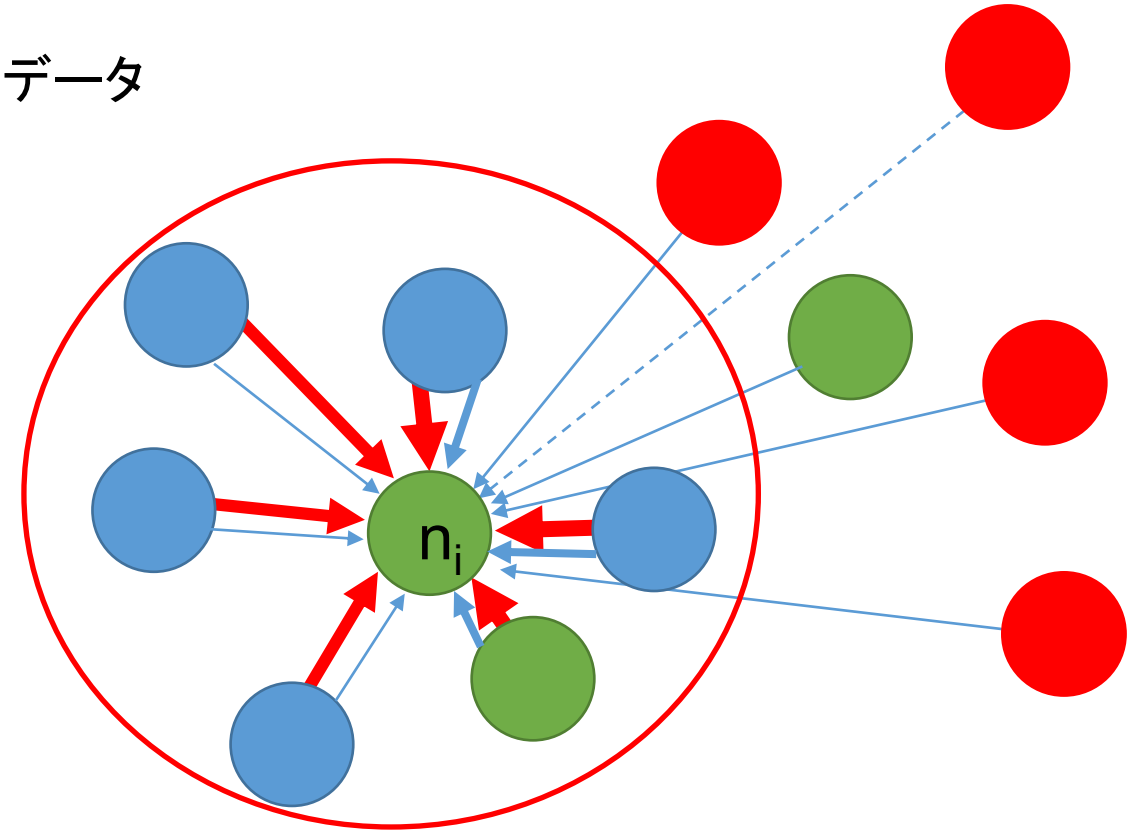
  2. 動的クランピング：ラベルありデータへの処理の拡張

アルゴリズム：局所的伝播

ラベルありデータ

ラベルなしデータ

top-k



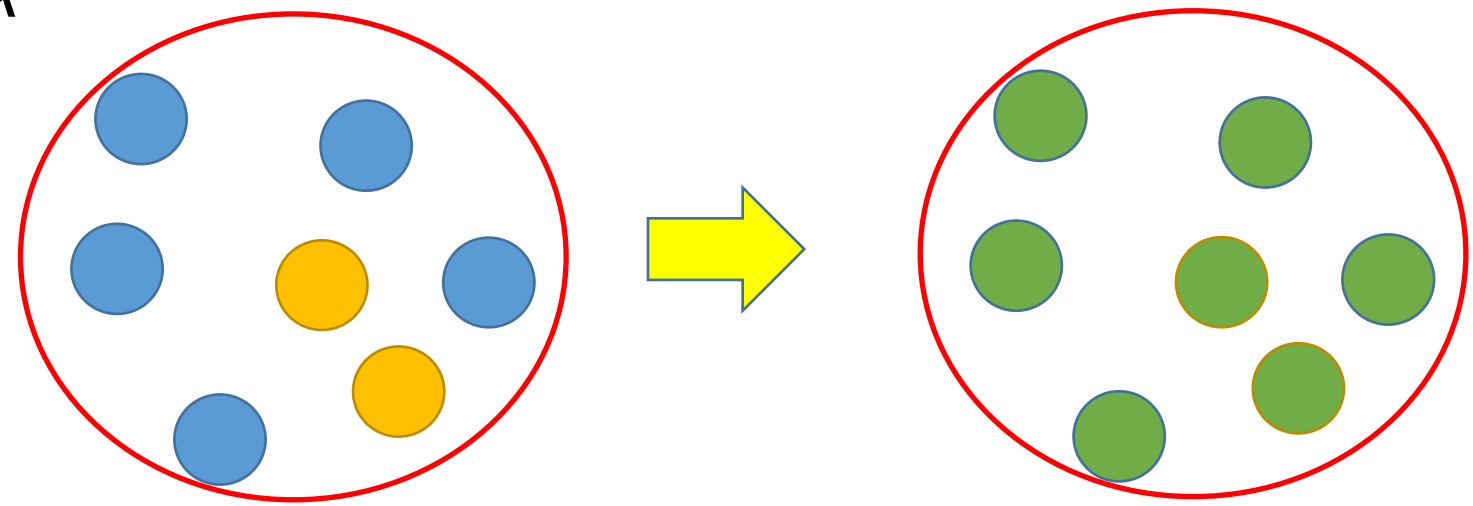
赤い矢印で表す類似データのラベル値を既存手法に追加して伝播

アルゴリズム：動的クランピング

n回目の反復処理時でのラベル値

平均値で更新した後のラベルの値

top-k



既存手法：
ラベルありデータのラベルを初期化
毎回同じ値に戻す（静的）

提案手法：
上位k個のデータのラベル値の平均。
伝播の反復処理ごとにこの値は変わる（動的）

実験

- 本発表ではWikipedia記事を対象にした評価結果を報告
- 精度の安定性はワークショップ論文（WII’18）で報告済
  T. Miyazaki and Y. Sumikawa, Label Propagation using Amendable Clamping, WII’18
  - SIAM2007テキストマイニングコンペのデータを使用
    - 22カテゴリ
    - 複数ラベルを持つ記事は約1万件
  - ラベル欠損が無い場合：
    - SVMが一番。マイクロ平均F値は約71%程度。
    - LPACは約70%
  - ラベル欠損が30%以上：LPACが一番良い。

実験｜データセットの統計情報

情報源	Wikipediaカテゴリ（Natural disasters）
教師データ	1,347
テストデータ	25
記事の長さの平均	7,878
カテゴリ数	6

データセットの詳細情報

- 対象のデータセット：英語版Wikipedia

- 対象のカテゴリ：Natural disastersとそのサブカテゴリ
  1. Avalanches（雪崩）：28
  2. Floods（洪水）：328
  3. Tornadoes（竜巻）：57
  4. Earthquakes（地震）：772
  5. Landslides（土砂崩れ）：158

- テストデータ：上記5カテゴリから5個ずつサンプリング

テストデータの作成手順

- ラベル欠損状況は人手による確認
  - 機械学習関係の博士号を持つ研究者3人による検証
- 25個のデータ全てに対して手動で多クラス分類を依頼
- この結果を評価に利用

実験｜比較対象

1. SVM

2. ランダムフォレスト（RF）

3. 動的LP（DLP）
   ラベル間の共起性を考慮した伝播を行うLPの最新手法
   Wang, B., Tsotsos, J.: *Dynamic label propagation for semi-supe-rvised multi-class multi-label classification*. Pattern Recognition 52, 75 – 84 (2016)

4. LPAC
   提案手法
   K=5

- 特徴ベクトルはLDA(t=8)を適用した結果の分布とした。

実験｜評価項目

RQ1. どの程度の良い精度（マイクロ平均F値）を得られる？
　　自動的にラベル付与できる程度の結果が得られている？

RQ2. 手動でのラベル付けの際に有効な結果を提示できる？
　　この結果を踏まえてラベル付けする時に労力を低減できるのか？

RQ1｜マイクロ平均F値

アルゴリズム	スコア
SVM	69.0%
RF	29.9%
DLP	40.8%
LPAC	78.9%

- 従来手法より良い結果が得られている。

RQ2｜手動ラベリングに有効？

各アルゴリズムが誤って付与したラベル（負例）と正しいが
付与していなかったラベル（欠損）の数を調査。

	SVM	LPAC
負例	0	11
欠損	26	11
合計	26	22

ボランティア協力者にLPACの結果を確認した
→ 負例は漏れなく排除できていた
→ 欠損数の少なさの観点と併せるとLPACの方が良い

本実験の限界

1. 他のデータセットでの結果は保証しない。
   - 違うカテゴリでの結果は予想できない。
   - 例：人物や組織のみのデータセットでの結果は不明。

2. 自然災害には他の種類もあるが網羅できていない。
   - 例えば「津波」も自然災害だが本実験では含まれていない。
   - 「地震」「津波」のような因果関係を含むデータは未評価

   → Wikipediaカテゴリシステムの到達性を向上させて欲しい。

まとめ

- 多クラス分類器(LPAC)を提案した。
  1. 局所的伝播：上位k個の値をより多く伝播する。
  2. 動的クランピング：伝播ごとに訓練データの値を更新する。

- 実際のラベル欠損を含むデータセットで評価した。

- 今後の課題
  1. ラベル間の共起性を考慮したラベル値の伝播法の提案。
  2. 負例も含むデータセットでも良い精度を出せるアルゴリズムの提案。