Vol.2011-NL-203 No.1
2011/9/16

教師データ間距離学習を利用した新語義用例の検出

佐々木　稔[*1]　新納浩幸[*1]

本稿では，データの一部にラベルが割り当てられた集合に対して，ラベル情報も考慮した外れ値の検出手法を提案し，用例集合から新語義として使用した用例候補の検出を行う．提案手法の有効性を評価するために，人工的に生成したデータによる外れ値検出を行う実験と Semeval-2010 日本語 WSD タスクのデータによる新語義用例検出を行う実験を行った結果，提案手法は外れ値の検出件数，および，F 値で LOF，One-Class SVM を上回る検出結果となり，密度に基づく新語義検出において，教師データの利用が有効であることが分かった．また，多くの用例について学習後に LOF 値の順位が上がり，距離学習による密度変化が新語義検出に有効であることが分かった．

Detection of Peculiar Examples using Distance Metric Learning from Labeled Example pairs

Minoru Sasaki[*1] and Hiroyuki Shinnou[*1]

In this paper, we propose a new peculiar example detection method using distance metric learning from labeled example pairs. To evaluate the efficiency of the proposed method, we make experiments on artificial dataset and Semeval-2010 Japanese WSD task dataset. As the results of these experiments, we found that it is effective for density-based outlier detection to use distance metric learning from the label information of training data.

1. はじめに

我々が生活の中で情報を収集する際に，知っている単語が知らない意味で使われた用例に接することも少なくない．その中には，国語辞典を調べても存在しない意味で使われたものも存在する．我々は，ある単語についての用例集合から，その単語の語義が新語義（辞書に未記載の語義）となっている用例を検出する研究に取り組んでいる．

新語義として使用した用例を検出するためのアプローチのひとつとして，データマイニング分野で使われる外れ値検出手法の適用が挙げられる．これは，用例集合の中で新語義の用例が特殊な単語の使い方をしていると考え，用例集合の中から外れ値を抽出することでそれが新語義の用例であると判定する手法である．ただし，ここで使われる外れ値検出手法はラベル情報を利用しない教師なし検出手法で，データの密度を利用して検出が行われる．そのため，新語義の検出で利用可能な教師データの語義情報を利用できず，語義情報も考慮した用例間の関連性を捉えることができない問題がある．

そこで本稿では，データの一部にラベルが割り当てられた集合に対して，ラベル情報も考慮した外れ値の検出手法を提案し，用例集合から新語義として使用した用例候補の検出を行う．この手法はラベル付きデータに対し，距離学習手法のひとつである Large Margin Nearest Neighbor(LMNN)[1] を利用して同じラベルを持つデータは近くに集め，異なるラベルを持つデータは遠くに移動することにより，ラベル情報を考慮したデータの分布を求める．この距離学習を行ったデータ集合に，外れ値の指標である Local Outlier Factor(LOF)[2] を利用することで外れ値候補を抽出する．

提案手法の有効性を評価するために，人工的に生成したデータによる外れ値検出を行う実験と Semeval-2010 日本語 WSD タスク[3] のデータによる新語義用例検出を行う実験を行う．Semeval-2010 日本語 WSD タスクのデータは本来語義識別を行うための評価データであるが，データの中に新語義の用例も含まれ，新語義用例検出の評価も行うことが可能である．提案手法と従来の外れ値検出手法である LOF，One-Class SVM を利用した手法による実験結果を比較し，外れ値検出における教師データの有効性を確認する．

2. 外れ値検出手法

従来の外れ値検出手法は，距離に基づく手法，確率モデルに基づく手法，データの密度に基づく手法など，数多く提案されている．本稿では，密度に基づく手法である LOF と，機械学習手法の Support Vector Machine(SVM) を用いた One-Class SVM について概要を説明する．

2.1 Local Outlier Factor(LOF)

LOF は，データ集合の各データに対してその近傍を考慮し，その密度を計算することで

*1 現在，茨城大学工学部情報工学科
Presently with Department of Computer and Information Science, Facalty of Engineering, Ibaraki University

得られる外れ値の度合いで，この値を比較することによりデータが外れ値かどうかを評価することが可能となる．

任意の正の整数 k に対して，データ x の k-distance(x) はデータ集合 D にある他のデータ $y \in D$ との距離 $d(x,y)$ を利用して，以下の 2 つの条件により定義される．

(1) 少なくとも k 個のデータ $y' \in D \setminus \{x\}$ に対して $d(x,y') \leq d(x,y)$ が成り立つ．

(2) 高々 $k-1$ 個のデータ $y' \in D \setminus \{x\}$ に対して $d(x,y') < d(x,y)$ が成り立つ．

これらの条件は複雑に見えるが，k-distance(x) の値はデータ x から k 番目のデータまでの距離となる．しかし，k 番目に同じ距離のデータが複数存在する場合は，この後の処理で求める k 個のデータ集合が一意に定まらないため，上記の条件を持つ定義となっている．

次に，データ x に近い k 個のデータ集合 $N_k(x)$ を求める．

$$N_k(x) = \{y \in D \setminus \{x\} | d(x,y) \leq k\text{-distance}(x)\} \tag{1}$$

k-distance(x) と $N_k(x)$ を利用して，データ x と y の距離関係を求める rearch-dist$_k(x,y)$，データ x の密度を表す $lrd_k(x)$ を以下の式で求める．

$$\text{rearch-dist}_k(x,y) = \max\{d(x,y), k\text{-distance}(y)\} \tag{2}$$

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} \text{rearch-dist}_k(x,y)} \tag{3}$$

これらの式を利用して，データ x が外れ値である度合いを表す $LOF(x)$ を以下の式と定義し，この値が大きいデータほど外れ値である可能性が高くなると推定することができる．

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)} \tag{4}$$

このとき，LOF にはパラメータとして k の値を決める必要がある．本稿では $k = 4$ として実験を行った．

2.2 One-Class SVM

One-Class SVM は，データ集合に対して高密度な領域を推定する教師なし学習アルゴリズムである．入力データ集合のラベルをすべて $+1$，原点のラベルを -1 と設定し，データ集合をカーネルに対応する非線形写像を用いて高次元空間上に射影し，その中で原点からラベル間のマージンが最大となる分離平面を求める．このとき，ソフトマージンを利用することで，外れ値である原点と同じ -1 のラベルと判定されたデータを求めることができる．このようにして求めたデータが外れ値として判定される．

この手法を利用する場合は，高次元空間に射影するためのカーネル関数の選択や識別誤りをどの程度許容するかを示すパラメータの設定によって，外れ値の識別結果に大きな影響を及ぼす．本稿における実験では，One-Class SVM のプログラムとして libsvm[⋆1] を用いた．また，カーネル関数は線形カーネルを選択し，識別誤りを許容するパラメータ n を $n = 0.02$ と設定して実験を行った．

3. 提 案 手 法

LOF は本来教師なし外れ値検出手法で，特徴空間内に存在するデータの密度を利用して，外れ値としての度合いを計算することで外れ値の候補を求めるものである．一般的に，外れ値と正常値を明確に定義することができないため，外れ値検出は教師あり手法を適用することができず，教師なしの枠組みにならざるをえない．そのため，LOF を新語義用例の検出に適用する場合は，語義ラベルを除いたデータだけしか使うことができず，教師データが外れ値として検出される可能性もある．

しかし，新語義用例検出においては，外れ値は辞書に記載されていない語義という明確な定義があり，正常値についても語義ラベルの付与されたデータが与えられていることで明確な定義が存在する．そのため，正常値については各語義ごとにクラスタができ，各クラスタの濃度が高くなるほど外れ値として検出されにくくなる．

そこで，教師データに対してラベルに応じてデータの移動を行う距離学習を利用し，同じラベルを持つデータは近くに集め，異なるラベルを持つデータは遠くに移動することを考える．距離学習を適用することにより，各語義について密度の大きいクラスタを作ることや密度の小さい語義間の境界を作ることが可能となる．このデータとテストデータに対して，密度に基づく外れ値検出手法を適用することにより，新語義用例をより効果的に検出可能になると考えられる．以上のことから，新しい新語義抽出手法として，教師データの距離学習手法として LMNN を利用し，LOF で新語義用例候補を抽出する手法を提案し，その有効性を検証する．

3.1 LMNN による距離学習

LMNN は，データそのものをラベルに応じて移動させることで最適なデータの位置関係を求める手法である．図 1 に示すように，データ $\mathbf{x}_i$ に近い指定した数の同じラベルのデータは近くに移動し，異なるラベルのデータはマージンが最大となるように移動する．このとき，近傍に存在するデータを表すフラグ行列 η を定義し，データ $\mathbf{x}_j$ が $\mathbf{x}_i$ の近傍にある場合に $\eta_{ij} = 1$，近傍にない場合は $\eta_{ij} = 0$ とする．このとき，目的関数となるコスト関数は

⋆1 http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Vol.2011-NL-203 No.1
2011/9/16

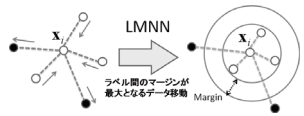

図 1 Large Margin Nearest Neighbor
Fig. 1 Large Margin Nearest Neighbor

以下のように定義され，この関数を最小とする変換行列 $\mathbf{A}$ を求める．

$$\epsilon(\mathbf{A}) = \sum_{ij} \eta_{ij} ||\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j||^2 + c \sum_{ijl} \eta_{ij}(1-\eta_{il})[1+||\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j||^2 - ||\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_l||^2]_+ \quad (5)$$

このコスト関数の第 1 項は同じラベルについての距離関係を表し，第 2 項は異なるラベルについての距離関係を表している．この関数を半正定値計画問題として最適解を求める．

4. 実　　験

本節では，提案手法の有効性を評価するために，人工的に生成したデータ集合と現実のデータ集合として Semeval-2010 日本語 WSD タスクで使用した集合の 2 種類を利用して，外れ値検出実験を行う．

4.1 人工データを利用した実験

実験で使用する人工データは，5 次元の正規分布を持つ 3 つのモデルからそれぞれ 200 個のデータを生成したものと外れ値として生成した 20 個のデータの合計 620 個のデータから成り立つ．5 次元正規分布モデルの各次元は独立であり，平均と分散は 0 以上 100 以下のランダムな値を設定し，各モデルからデータの生成を行う．また，外れ値は正規分布モデルから生成された 600 個のデータの最大値と最小値を求めて，その範囲内の値で一様分布となるように 5 次元のデータを 20 個生成する．

次に，正規分布モデルから生成されたデータ集合に対して，提案手法で適用する距離学習を行うために，学習データとテストデータの 2 つ集合に分割する．提案手法では，各モデルから生成された 200 個のデータのうち 20 個を教師データとして使用し，この分割により得られた合計 60 個の教師データを利用して，LMNN を用いた距離学習を適用することで教師データの分布をラベルに応じて更新する．

以上の方法により生成したデータに対して，提案手法と従来手法である LOF，One-Class SVM(OCS) により外れ値検出の実験を行った結果を表 1 に示す．ここで，LOF による手法と提案手法では，LOF 値の大きな上位 20 個のデータを外れ値の候補として取り出すこととする．また，One-Class SVM による手法では，30 個のデータを外れ値の候補として抽出する．

表 1 人工データに対する実験結果
Table 1 Experimental results on artificial dataset

手法	抽出数	正解数	F 値
LOF	20	12	0.600
OCS	30	12	0.480
提案手法	20	15	0.750

表 2 新語義検出実験の対象単語
Table 2 Target words used in the experiment

品詞	単語
名詞	相手，意味，関係，技術，経済，現場，子供，時間，市場，社会，情報，手，電話，場合，はじめ，場所，一，文化，ほか，前，もの
動詞	会う，あげる，与える，生きる，入れる，教える，考える，勧める，する，出す，立つ，出る，とる，乗る，始める，開く，見える，認める，見る，持つ，求める，やる
形容詞	大きい，高い，強い，早い，良い

4.2 Semeval2010 日本語タスクのデータによる実験

4.2.1 デ　ー　タ

本実験で使用するデータは，Semeval-2010 日本語 WSD タスクで課題として公開されたデータを利用する．これは指定された 50 個の対象単語について，各単語が含まれる文章が教師データ，テストデータとして各 50 個用意されている．その中に存在する対象単語には語義ラベルが付与され，教師データはそのラベル情報を元にモデル構築などの知識として利用し，テストデータのラベルはそのモデルの評価に利用する．

本稿における実験では，対象となる 50 単語のうち「可能」と「入る」は学習データ内に新語義のラベルが付与された用例があるため対象から外し，残りの 48 単語を実験の対象とした．対象となる単語を表 2 に示す．この中で，新語義の用例は「意味」で 1 用例,「手」で 3 用例,「前」で 7 用例,「求める」で 1 用例,「あげる」で 2 用例,「始める」で 2 用例の合計 16 用例存在する．これらの用例を検出することが目的である．

4.2.2 実 験 方 法

まず，実験対象となる各単語の教師データとテストデータとして与えられた合計 100 用例に対して特徴抽出を行う．特徴としては対象単語の前後 2 単語の品詞情報，分類番号を利用し，各用例について特徴ベクトルを作成する．LOF と One-Class SVM では得られた 100 個の特徴ベクトルを利用して外れ値候補の検出を行う．提案手法では，教師データの特

表 3 Semeval-2010 日本語 WSD タスクのデータに対する実験結果
Table 3 Experimental results on Semeval-2010 Japanese WSD task dataset

手法	抽出数	正解数	F 値
LOF	960	3	0.006
OCS	1150	3	0.005
提案手法	960	5	0.012

表 4 各新語義用例における距離学習の効果
Table 4 Effect of distance metric learning for each peculiar example

単語	用例番号	学習前順位	学習後順位
意味	31	84	12
手	32	14	14
	33	20	18
	34	37	19
前	34	54	64
	35	51	60
	36	53	63
	38	9	17
	46	50	59
	47	59	66
	48	98	95
求める	31	67	46
あげる	40	89	65
	41	93	55
始める	48	59	42
	49	57	38

徴ベクトルと語義ラベルに対して，LMNN を用いた距離学習を適用することで教師データの分布をラベルに応じて更新する．そこで得られた特徴ベクトルとテストデータの特徴ベクトルからなるデータ集合に，LOF を利用して外れ値候補の検出を行う．

実験の結果を表 3 に示す．LOF と 提案手法については，LOF 値の上位 20 個を新語義用例候補として抽出を行った．

5. 考　　察

表 1 および，表 3 より，人工データ，Semeval-2010 日本語 WSD タスクのデータの両方において，提案手法は外れ値の検出件数，および，F 値で LOF，One-Class SVM の結果を上回っている．このことより，密度に基づく新語義検出において，教師データの利用が有効であることが分かる．特に，人工データにおいてはその一部を教師データとして利用することで，外れ値をより多く検出することが可能となっている．

表 4 は，距離学習前後における LOF 値の順位変化を表している．多くの用例について学習後に LOF 値の順位が上がり，距離学習による密度変化が新語義検出に有効であることが分かる．単語「前」については，すべてが「午前」を省略した形であり，前後に漢数字が共起した用例である．順位が唯一高い用例番号 38 は「＝前８・三十」と，この用例のみアラビア数字が共起している．教師データには「十年前」という記述があることから，漢数字の「十」が順位を下げる要因になっているかと考えられる．

6. おわりに

本論文では，対象単語の用例集合の一部に語義ラベルが割り当てられた集合に対して，ラベル情報も考慮した新語義用例の検出手法を提案した．提案手法の有効性を評価するために，人工的に生成したデータによる外れ値検出を行う実験と Semeval-2010 日本語 WSD タスクのデータによる新語義用例検出を行う実験を行った結果，提案手法は外れ値の検出件数，および，F 値で LOF，One-Class SVM を上回る検出結果となり，密度に基づく新語義検出において，教師データの利用が有効であることが分かった．また，多くの用例について学習後に LOF 値の順位が上がり，距離学習による密度変化が新語義検出に有効であることが分かった．

今回の実験では，距離学習を行う訓練データの数が一定であったため，それを変化させた時の検出効果の変化，テストデータも距離学習に使うための方法について工夫することが今後の課題となる．

参　考　文　献

1) Weinberger, K.Q. and Saul, L.K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, *The Journal of Machine Learning Research*, Vol.10, pp. 207–244 (2009).
2) Breunig, M.M., Kriegel, H.-P., Ng, R.T. and Sander, J.: LOF: identifying density-based local outliers, *SIGMOD Rec.*, Vol.29, pp.93–104 (2000).
3) Okumura, M., Shirai, K., Komiya, K. and Yokono, H.: SemEval-2010 Task: Japanese WSD, *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, Association for Computational Linguistics, pp.69–74 (2010).