外れ値検出手法を利用した新語義の検出

新納浩幸
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1　はじめに

本論文では対象単語の用例集合から，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手法を提案する．

新語義の検出は語義曖昧性解消の問題に対する訓練データを作成したり，辞書を構築する際に有用である．また新語義の検出は意味解析の精度を向上させる [5]．また新語義の用例はしばしば書き誤りとなっているので，誤り検出としても利用できる．新語義検出は一般に Word Sense Disambiguation (WSD) の一種として行う方法，新語義の用例をクラスターとして集める Word Sense Induction (WSI) のアプローチで行う方法 [3]，及び新語義の用例を用例集合中の外れ値とみなし，外れ値検出の手法を用いる方法 [5] がある．ここでは外れ値検出の手法のアプローチを取る．ただしデータマイニングで用いられる外れ値検出の手法は教師なしであるが，本タスクの場合，少量の用例に語義のラベルが付いているという教師付きの枠組みで行う方が自然であり，ここでは教師付き外れ値検出の手法を提案する．

提案手法は 2 つの検出手法からなる．第 1 の手法は代表的な外れ値検出手法である Local Outlier Factor (LOF)[9] を教師付きの枠組みに拡張したものである．第 2 の手法は，対象単語の用例（データ）の生成モデルを用いたものである．一般に外れ値検出はデータの生成モデルを構築することで解決できる．提案手法では第 1 の手法と第 2 の手法の出力の積を取ることで，最終の出力を行う．

提案手法の有効性を確認するために，SemEval-2 の Japanese WSD タスク [8] のデータを利用した．従来の外れ値検出の手法と比較することで提案手法の有効性を示す．実験を通して，外れ値検出に教師データを利用する効果も確認する．また SVM による WSD の信頼度を利用した外れ値検出も行い，WSD の延長では新語義の検出が困難であることも示す．

2　従来の新語義検出手法

2.1　WSD の信頼度の利用

WSD は語義を識別するので，WSD システムを利用すれば新語義を検出できると考えるのは自然である．

WSD の対象単語 w の語義のクラスを C とする．関数 $f(x,c)$ はある WSD システムが出力する用例 x 中の w の語義が $c \in C$ となる信頼度とする．この WSD システムは $\arg\max_{c \in C} f(x,c)$ により語義を識別する．

新語義の検出はある閾値 θ を定め，

$$\forall c \in C \quad f(x,c) < \theta \tag{1}$$

のときに x を新語義の用例と判定することで，新語義を検出できる．

ただし WSD は識別のタスクであり，一般に WSD システムは SVM のような識別モデルをもとに構築される．そのためシステムは語義の識別精度が上がるように最適化されており，$f(x,c_i)$ の値は $f(x,c_j)$ との相対的なものであり式 (1) により新語義が検出できる保証はない．

2.2　WSI による検出

従来，新語義の検出は Word Sense Induction (WSI) というタスクの一部として行われてきた．これは本質的には対象単語の用例を語義に基づいてクラスタリングするタスクである [3]．用例集合中に新語義の用例があれば，それらも語義のクラスターとして出現するために新語義の検出として利用できる．

ただし陽に新語義を検出するには，得られたクラスターに語義のラベルを付与する必要がある．Shirai は辞書に記述された語義の定義を利用して，得られたクラスターに語義のラベルを付けることで新語義を検出しようとしている [6]．また Sugiyama は既存語義の用例を種用例として，用例集合を半教師なしクラスタリングによりクラスタリングした．種用例のないクラスターが新語義のクラスターとなる．ただしどちらもクラスタリング自体の精度が悪く，新語義の検出までには至っていない．

本来，クラスターに語義のラベルを付けるためには，語義のラベル集合が必要である．語義のラベル集合を定めた場合に，WSI と WSD との違いはほとんどなくなる．WSD を行う前に教師なし学習であるクラスタリングを行うアプローチが，新語義の検出に有効かどうかは不明である．また用例を語義に基づいてクラスタリングする場合，クラスターの数の決め方が大きな問題になる [1]．

2.3　外れ値検出による検出

新語義の用例を用例集合内の外れ値と見なし，外れ値検出の手法を利用して新語義を検出するアプローチがある．

Erk は外れ値検出手法の最近傍法を利用して新語義の検出を試みた [5]．対象単語 w の語義が付与された用例集を D とし，用例 x の外れ値の度合い $out(x)$ を式 (2) で測り，この値が 1 以上の x を新語義の用例とした．ここで $d(x,y)$ は用例 x と用例 y 間の距離である．

$$out(x) = \frac{\min_{y \in D} d(x,y)}{\min_{z \in D} d(y,z)} \tag{2}$$

ただし最近傍法が妥当な精度を出すには，大量の訓練データを必要とするという問題がある．

3 外れ値検出手法

データマイニング分野の外れ値検出手法は非常に多岐にわたるが，その多くは変化点検出の手法に位置づけられる [11]．つまり時系列的にデータが生起するオンラインでのタスクに対する手法が中心である．新語義検出のようなバッチ的なタスクに対する手法としては，密度ベースの手法，One Class SVM，生成モデルによる手法が代表的な手法である．ここではこの 3 つの手法を，本論文の提案手法との比較手法とする．

3.1 密度ベースの手法

LOF は，データの近傍の密度を利用することで，そのデータの外れ値の度合いを測り，その値によって外れ値を検出する．

LOF におけるデータ $x \in D$ における外れ値の度合いを $LOF(x)$ と表記する．ここで D はデータ全体の集合である．$LOF(x)$ を定義するために，いくつかの式を定義しておく．まず $kdist(x)$ は x に対する k 距離と呼ばれる値で，データ x からの k 番目に近いデータまでの距離である．

次に $kdist(x)$ を利用して，$N_k(x)$，$rd_k(x,y)$ 及び $lrd_k(x)$ を以下のように定義する．

$$N_k(x) = \{y \in D \setminus \{x\} | d(x,y) \leq kdist(x)\}$$

$$rd_k(x,y) = \max\{d(x,y), kdist(y)\}$$

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} rd_k(x,y)}.$$

これらの式を用いて，$LOF(x)$ は以下で定義される．

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

また LOF ではパラメータとして k が存在する．本論文では $k = 4$ としている．

3.2 One Class SVM

One Class SVM は ν–SVM[2] を利用した外れ値検出の手法である．すべてのデータは $+1$ のクラスに属し，原点のみが -1 のクラスに属するとして，ν–SVM を使って 2 つのクラスを分離する超平面を求める．原点はすべての点に対して類似度が 0 となるために，外れ値とみなせる．また ν–SVM はソフトマージンを利用するので，-1 のクラス側に属するデータを外れ値と判定する．

One Class SVM を利用する際には，用いるカーネル関数やどの程度のマージンの誤りを認めるかのパラメータの設定が結果に大きく作用する．本論文の実験では One Class SVM のプログラムとして libsvm [1] を用いた．カーネルは線形カーネルを利用し，マージンの誤りはパラメータ n に対応するが，$n = 0.02$ で固定した．

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

3.3 生成モデルによる手法

データ x の生起する確率モデル $P(x)$ を生成モデルと呼ぶ．一般に潜在変数 z_i を導入し，ある確率モデル $P_i(x)$ の混合分布により $P(x)$ をモデル化する．

$$P(x) = \sum_i z_i P_i(x) \quad s.t. \quad \sum_i z_i = 1$$

モデル化の後に，与えられたデータから EM 法などを利用して，z_i と $P_i(x)$ のパラメータを推定することで $P(x)$ を構成する．

データ x の外れ値の度合いとしては $-\log P(x)$ が用いられる．この値が大きいほど外れ値と見なせる．

4 提案手法

4.1 教師付き外れ値検出

一般に外れ値検出のタスクでは外れ値の定義が不可能である[2]．これは外れ値にラベルをつける意味がないことを示している．なぜなら仮にあるデータが外れ値であり，その外れ値にラベルをつけることができたとしても，他の外れ値がそのラベル付きの外れ値と類似している保証がないからである．また検出元となるデータ集合は，ほぼすべて正常値である．仮にデータにラベルをつけるとすれば，正常値のラベルだけになり，教師データに意味はない．これらのことから外れ値検出の手法は教師なしの枠組みにならざるおえない．

しかし新語義を外れ値と見なした新語義検出のタスクの場合，一般の外れ値検出とは異なった 2 つの特徴がある．1 つは外れ値の定義が明確である点である．ここでの外れ値は新語義の用例であるが，新語義とは辞書に記載されていない語義である，というように明確に定義できる．もう 1 つは正常値のデータは語義のクラスターに分割されるという点である．しかもクラスターの数も明確である．一方，通常の外れ値検出では正常値の集合がクラスターに分割されるのか，されるとしてもいくつのクラスターに分割されるのかは不明である．

ここではこれらの特徴を利用して外れ値検出を行う．つまり，検出元となる対象単語の用例集の一部に，対象単語の語義のラベルを付与し，その設定のもとで外れ値検出を行う．

4.2 教師付き LOF

教師データを LOF で利用するには単純に教師データをテストデータに加えればよい．しかしその場合，教師データからも外れ値が検出される可能性がある．

ここでは教師データを $k+1$ 倍してからテストデータに加えてデータセットを作り，そのデータセットに対して LOF を適用する．ただし k は LOF における $kdist$ で使われる k である．

LOF の場合，訓練データ x を $k+1$ 倍すると $kdist(x) = 0$ となり，訓練データ x が外れ値として検出されることはなくなる．さらにテストデータ y と訓練データ x との距離が小さいと，その訓練データ x は $k+1$ 個存在するために，テストデータ y の密度も高まり，外れ値としては検出されなくなる．

[2] もしも定義できるのであれば，その定義にあったデータを取り出せばよいだけなので，タスクとしての意味はなくなる．

4.3 教師付き生成モデル

対象単語 w の用例 x に対する生成モデル $P(x)$ を構成する．w の語義を $z_i (i = 1 \sim K)$ としたとき，全確率の公式から，以下が成立する．

$$P(x) = \sum_{i=1}^{K} P(z_i)P(x|z_i)$$

w の訓練データが N 個あり，その中で語義 z_i のデータが n_i 個あるとする．また x は以下のような素性リストで表現されており，

$$x = \{f_1, f_2, \cdots, f_m\}$$

訓練データの中の語義が z_i となっているデータの中で，f_j が出現した個数を $n(z_i, f_j)$ と書くことにする．このとき，MAP 推定でスムージングを行い，以下の推定式が得られる [10].

$$P(z_i) = \frac{n_i + 1}{N + K} \tag{3}$$

$$P(x|z_i) = \prod_{j=1}^{m} P(f_j|z_i) = \prod_{j=1}^{m} \frac{n(z_i, f_j) + 1}{n_i + 2} \tag{4}$$

以上より $P(x)$ の値が求まる．外れ値の度合いは $-\log P(x)$ で測れるので，この値の大きなものを外れ値の候補とする．

5 実験

実験データとして SemEval-2 の Japanese WSD タスクのデータを用いる．Japanese WSD の語義識別の対象の単語は 50 単語である．この中で「可能」「入る」は教師データ内に新語義の用例があるので，それらを外して，残り 48 単語を実験対象とした．

新語義は「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,「はじめる」で 2 用例の計 16 用例存在する．これらが検出の正解となる．

実験の結果を表 1 に示す.

表 1: SemEval-2 データに対する実験結果

手法	抽出数	正解数	F 値
LOF	240	2	0.0156
OCS	1228	3	0.0048
LOF+OCS	83	0	0.0000
NN	111	0	0.0000
S-LOF	240	5	0.0391
G-model	320	5	0.0298
本手法	28	2	0.0909

LOF では LOF 値の大きなもの上位 5 個を取り出すことにする．OCS は One Class SVM の意味である．LOF+OCS は LOF の出力と OCS の出力の積をとったものである [4]．NN は [5] で用いられた最近傍法であり式 (2) が 1.0 以上のものを取り出している．S-LOF は本論文で提案した教師付き LOF を指す．また G-model は本論文で説明した生成モデルによるものである．ただし各単語に対して，$-\log P(x)$ の値を正規化し，正規化した値が 1.1 以上のものを取り出すことにする．表 1 から本手法の有効性は明かである．

6 考察

6.1 WSD による新語義検出

本実験と同じ素性を使い SVM を学習し，SemEval-2 Japanese WSD タスクのテストデータ 50 単語全てを対象に語義の曖昧性解消を行ったところ，平均正解率は 0.7664 であった．上記タスクの参加システム中最高の正解率は RALI-2 の 0.7636 であり [8]，ここで学習できた SVM は十分能力が高いことがわかる．

ここでの SVM の学習には LIBSVM[3] を用いたが，そこでは`-b` のオプションで識別の確信度を求めることができる．このオプションを用いて，語義の数が K の場合，$1.1/K$ よりも小さな確信度の場合に，その用例を新語義の用例とすることで新語義の検出を試みた．結果，検出数は 56 正解数は 1 となり，F 値は 0.0278 であった．この値は表 1 で示された外れ値検出手法と比較すると，それほど悪いとも言えないが，WSD システム単独では新語義の検出が困難であることがわかる．

6.2 未出現語義を含めた評価

SemEval-2 Japanese WSD タスクでは，訓練データ中には現れないが，テストデータには出現する語義が存在する．このような訓練データ中の未出現語義は，新語義と見なすこともできる．このような用例は「あう」で 1 用例,「すすめる」で 1 用例,「出す」で 3 用例,「立つ」で 1 用例,「とる」で 3 用例,「ひとつ」で 1 用例,「見る」で 6 用例,「持つ」で 1 用例,「大きい」で 2 用例,「与える」で 1 用例の合計 20 用例存在する．これらも新語義の用例と見なした場合の検出結果を表 2 に示す．この場合でも本手法が最も F 値が高かった．また，先の WSD システムを用いた場合は，検出数は 56 正解数は 1 のままであり，F 値は 0.0217 と下がる．

表 2: 未出現語義を含めた評価

手法	抽出数	正解数	F 値
LOF	240	3	0.0217
OCS	1228	10	0.0158
LOF+OCS	83	0	0.0000
NN	111	3	0.0408
S-LOF	240	6	0.0435
G-model	320	8	0.0449
本手法	28	2	0.0625

6.3 誤検出・未検出の原因

本手法の誤検出の原因について述べる．1 つは固有表現や熟語内の単語である．例えば以下のような表現が検出されている．

(a) 未来科学 技術 共同研究センターの中の研究施設
(b) 昔話の「千代ごこ 出 やっせ」のように
(c) 中小零細企業の取材は数多く 手 がかかる割りに

固有表現や熟語内の単語に通常の意味があるとは考えづらく，新語義の検出という観点では，このような表現を抽出しても完全に誤りとは言えない．本来，新語義の検出するためには，固有表現や熟語を予め抽出しておくことが必要だと考える．

[3] `http://www.csie.ntu.edu.tw/ cjlin/libsvm/`

また誤検出のその他の原因は多様であるが，全体として，対象単語の直前や直後に自立語が現れる複合語の用法や動詞の連体形の用法などが目立った．

(d) わが国が最も重要な貿易 相手 国の一つ
(e) 人間性を疑ってしまう人とは男女 関係 なく，
(f) 夏休み等に行って来た時の経験＝古き 良き 時代を，

複合語が専門性の高い用語である場合は意味のある検出とも見なせるが，ここでは複合語を単なる名詞連続で認識しているために，専門用語との区別は付けられない．新語義の検出に関しては，熟語や固有表現と同様，専門用語も通常の表現とは，区別した方がよいと考える．

本手法の未検出の原因としては，突き詰めれば，用例間の距離の測定方法に帰着される．ある新語義の用例と他の正常値の用例との距離がある程度，離れていたとしても，正常値の用例間の距離も同程度は離れているという状況である．これは動詞や形容詞における検出では顕著である．この解決は語義識別の場合と同じであり，語義識別の精度向上の試みが本研究に応用できると考えている．

6.4 外れ値検出手法のバギング

外れ値検出手法は数多く提案されているが，その有効性はタスクに依存するため，新語義検出に適した外れ値検出手法が望まれる．ただし新たに外れ値検出手法を考案するのではなく，既存の手法を組み合わせる戦略も有効である．Lazavic は複数の外れ値検出の手法を適用して，それら出力結果を総合的に判断して最終的に外れ値候補を出力するという外れ値検出手法のバギング (Bagging) を提案した [7]．ここで提案した LOF と生成モデルの組み合わせも，外れ値検出手法のバギングの一種と考えられる．ここでは単純に出力の積により最終の出力を決めたが，重みを付けて判断するなどの工夫も考えられる．あるいは他の外れ値検出の手法の組み合わせることも有効であろう．表 2 からもわかるとおり，LOF の出力と生成モデルの出力はかなり異なる．単純に出力の和を取ると，検出数が多くなりすぎて F 値の評価は下がってしまうが，第 1 段目の候補としては取り出せているので，そこからの選別に工夫することで改善が可能だと考えている．ここらが今後の課題である．

7 おわりに

本論文では対象単語の用例集合から，その単語の語義が新語義となっている用例を検出する手法を提案した．基本的に新語義の用例を用例集合中の外れ値と考え，外れ値検出の手法を利用する．ただし従来の外れ値検出では教師なしの枠組みであるが，ここではタスクの性質を考え，教師付きの枠組みで行った．

提案手法は教師データを利用した LOF と生成モデルによる出力の積を取るものである．SemEval-2 の Japanese WSD タスクのデータを用いた実験により，提案手法の有効性を示した．また WSD システム単独では新語義の検出が困難であることも示した．

様々な外れ値検出手法を統合するバギングの手法を用いて，検出の精度を向上させることが今後の課題である．

参考文献

[1] Eneko Agirre and Aitor Soroa. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *SemEval-2007*, 2007.

[2] B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, Vol. 13, No. 7, pp. 1443–1471, 2001.

[3] Michael Denkowski. Survey of Techniques for Unsupervised Word Sense Induction, 2009.

[4] Hiroyuki Shinnou and Minoru Sasaki. Detection of Peculiar Examples using LOF and One Class SVM. In *LREC-2010*, 2010.

[5] Katrin Erk. Unknown word sense detection as outlier detection. In *NAACL-2006*, pp. 128–135, 2006.

[6] Kiyoaki Shirai and Makoto Nakamura. JAIST: Clustering and Classification Based Approaches for Japanese WSD. In *The 5th International Workshop on Semantic Evaluation*, pp. 379–382, 2010.

[7] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *The eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)*, pp. 157–166, 2005.

[8] Manabu Okumura and Kiyoaki Shirai and Kanako Komiya and Hikaru Yokono. SemEval-2010 Task: Japanese WSD. In *The 5th International Workshop on Semantic Evaluation*, pp. 69–74, 2010.

[9] Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *ACM SIGMOD 2000*, pp. 93–104, 2000.

[10] 高村大也．言語処理のための機械学習入門．コロナ社, 2010.

[11] 山西健司．データマイニングによる異常検知．共立出版, 2009.