教師付き外れ値検出による新語義の発見

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。

新語義の発見は語義識別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。ここでのアプローチの基本は、新語義の用例が用例集合中の外れ値になると考え、データマイニング分野の外れ値検出の手法を利用することである。ただし外れ値検出のタスクは教師なしの枠組みになるが、新語義検出という本タスクの性質を考慮すると、一部のデータ（用例）にラベル（対象単語の語義）が付与されているという枠組みで考える方が適切である。そのため本論文では一部のデータにラベルがついているという教師付きの枠組みで外れ値検出を行う。

提案手法は 2 つの検出手法からなる。第 1 の手法は従来の外れ値検出手法である Local Outlier Factor (LOF)[4] 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、教師データから語義識別の分類器を学習し、各データの語義を推定する。推定された語義のクラスターとデータとの距離関係から外れ値かどうかを判定する。提案手法では第 1 の手法により外れ値の候補を取り出し、第 2 の手法でその候補を選別する。

提案手法の有効性を確認するために、2 つの実験を行った。人工的に作ったデータに対するものと、SemEval-2 の Japanese WSD タスク [3] のデータに対するものである。SemEval-2 の Japanese WSD タスクは通常の語義識別のタスクであるが、識別する語義の対象に新語義を含めている点に大きな特徴がある。このためこのタスクの訓練データを教師データとして利用して、テストデータから新語義を検出するという設定で実験が行える。

2 つの実験を通して、外れ値検出に教師データを利用する効果が確認できた。今後の課題としてはパラメータの設定法がある。本手法ではパラメータが 3 つ存在し、これらの値が結果に大きく作用する。またタスクに応じて適切な値が異なる。このため適切な設定方法が必要である。

2　教師付き外れ値検出

2.1　従来の外れ値検出手法

外れ値検出の手法は多岐にわたるが、おおまかに分類するとデータの生成に確率モデルを用いるものと用いないものに分けられる [5]。確率モデルを用いた場合、データの生成確率が得られるので、その確率が低いデータを外れ値とすればよい。このアプローチでは、いかに適切な確率モデルを構築できるかが鍵となる。確率モデルを用いない手法としては LOF[4] と One Class SVM[1] が代表的である。

2.1.1　LOF

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、以下の条件を満たすデータ $o \in D$ との距離 $d(x,o)$ として定義される。

1. 少なくとも k 個のデータ $o' \in D \setminus \{x\}$ に対して $d(x,o') \leq d(x,o)$が成立する。
2. 高々$k-1$ 個のデータ $o' \in D \setminus \{x\}$ に対して $d(x,o') < d(x,o)$が成立する。

直感的には、上記のデータ o はデータ x からの k 番目に近いデータとなる。データ x から同じ距離を持つデータが複数存在する場合を考慮して、上記のようなテクニカルな定義になっている。

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 $k = 4$ を用いている。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

2.1.2 One Class SVM

One Class SVM は ν–SVM[1] 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。

2.2 外れ値検出と新語義検出

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。具体的には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。

2.3 語義識別問題としての新語義検出

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た場合、帰納学習の手法を利用して語義識別を行う分類器を学習することができる。この分類器の識別の信頼度を利用して新語義の検出を行える可能性がある。ただしこのような分類器の識別の信頼度を利用する方法では新語義の発見は困難である。この点を簡単に注記しておく。

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

[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。

基本的に帰納学習で得られる分類器は、入力されるデータが与えられたクラスのいずれかに属することを仮定しており、その仮定の下で識別精度を高めることを目指している。例えば、SVM では分離平面だけが問題であり、クラスターの構造を考慮しない。そのため図 1 のような状況では、データ a とデータ b の識別の信頼度は同じであるが、明らかに外れ値の度合いはデータ b の方が高い。

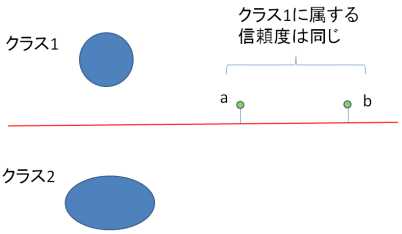

図 1: 識別の信頼度と外れ値の度合い

3 提案手法

ここでは外れ値（新語義の用例）を検出するために、2 つの手法を提案し、それらを組み合わせる。第 1 の手法は LOF を教師データを利用するように拡張したものであり、第 2 の手法は教師データから分類器を作成し、データのクラスを識別し、データと識別されたクラス間の距離関係から外れ値かどうかを判定するものである。第 1 の手法で外れ値の候補を取り出し、第 2 の手法でそれらを選別する。

3.1 教師データ付き 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。

3.2 クラス推定とクラスとの距離関係

教師付き LOF の場合、ラベル（語義）の種類による区別はない。何らかのラベルが付与されていれば、すべて正常値という扱いになる。ここでは教師データのラベルの種類を利用することを考える。

外れ値検出では、クラスタの分布がわかれば、外れ

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

値かどうかの判断は閾値の問題だけになる。例えばクラスタの分布が多次元正規分布であれば、マハラノビスの距離からデータとクラスタ間の距離が測れるので、それによって外れ値の識別が可能になる。しかし教師データを用いたとしてもクラスタの分布の推定は困難なことが多い。

ここではクラスタの分布を仮定せずに、データがそのクラスタに対して外れ値になるかどうかを判定する。

まず教師データからクラスを識別する分類器を学習する。データ x に対してその分類器を用いて、そのデータのクラス A を推定する。次に A の中でデータ x に最も近いデータ $y \in A$ を見つけ、x と y 間の距離 $d(x,y)$ と y と A の重心 $\bar{A}$ 間の距離 $d(y,\bar{A})$ を測る。これらの比 r を求めて、ある値 r_0 以上のものを外れ値を判断する。

$$r = \frac{d(x,y)}{d(y,\bar{A})}$$

4 実験

ここでは提案手法の有効性を確認するために、人工的なデータと現実のデータを用いる。現実のデータは SemEval-2 の Japanese WSD タスクで使われたデータである。

4.1 人工データによる実験

3 つの 5 次元正規分布のモデル[3]を作り、それぞれのモデルから 200 個のデータを生成する。各モデルから作られた 200 個のデータのうち 20 個を教師データする。また作成された 600 個の全データ内の最大値 max と最小値 min を求め、$[min, max]$ の範囲の一様分布から 5 次元の点を 20 個作成する。これが外れ値ある。以上よりクラス数は 3 個、教師データは 60 個、検出対象のデータは 560 個、うち外れ値は 20 個となる。

これらのデータに対して、LOF、One Class SVM (OCS)、それらの積を出力するもの (LOF+OCS)[2]、教師付き LOF (S-LOF)、本手法の結果を以下に示す。ただし LOF では LOF 値の大きなもの上位 20 個を取り出すことにする。また本手法を使う際には $r_0 = 1$ とした。

表 1: 人工データに対する実験結果

手法	抽出数	正解数	F 値
LOF	20	12	0.600
OCS	30	12	0.480
LOF+OCS	8	5	0.357
S-LOF	20	12	0.600
本手法	10	10	0.667

また識別の信頼度による外れ値の検出も試みた。今、クラスは A、B、C の 3 つあるので、A か A 以外、B か B 以外、C か C 以外を識別する SVM を 3 つ学習し、各 SVM の結果が A 以外、B 以外、C 以外となった場合に、そのデータを外れ値とすることにした。この場合、検出数は 174 、正解数は 3 となり、検出の F 値は 0.031 であった。なおここで学習された SVM は外れ値を除いたテストデータ 540 個に対する正解率は 100% であり、識別の精度がよくても外れ値の検出は困難であることがわかる。

[3]各次元は独立、各次元の分散は異なるモデルを利用した。平均と分散は 0 以上 100 以下の値からランダムに取り出した。

4.2 SemEval-2 Japanese WSD タスクのデータによる実験

SemEval-2 は語義曖昧性解消に関する評価型の国際会議であり、いくつかのタスクが設定されている。Japanese WSD はその中の 1 つである。通常の日本語の語義識別のタスクであるが、最も特徴的な点は、識別結果に新語義というカテゴリを含めている点である。つまりテストデータの中には設定された語義のどれでもないという答えがありえる。そのため、このタスクで用意された訓練データとテストデータを用いることで、教師付きの枠組みでの新語義の検出手法の評価が可能である。

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。各単語を以下に示す。

名詞　21 単語
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　22 単語
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　5 単語
大きい、高い、強い、早い、良い

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。

実験の結果を以下に示す。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。また本手法を使う際には $r_0 = 3$ とした。

表 2: 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	0	0.000
OCS	1150	3	0.005
LOF+OCS	83	0	0.000
S-LOF	240	3	0.023
本手法	36	2	0.077

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

5 考察

人工データに対する実験結果は以下の点を示している。

1. 教師なしの LOF や One Class SVM でも、ある程度の検出は可能である。
2. LOF に教師データを利用する効果は少ない。
3. 推定クラスとの距離を測る本手法のフィルターは有効である。

ただし SemEval-2 のデータに対する実験結果を見ると、(1) や (2) は逆になる。人工データはデータの生成が単純なモデルで表現できている。このような場合は教師なしの手法でもうまくいくが、SemEval-2 のデータのようにデータの生成が複雑、つまり正常値のクラスターが複雑な形状をしている場合は、教師なしの手法は有効に働かず、教師データを利用する効果が高い。実験では、教師データを利用することで抽出できなかった新語義も抽出できている。また LOF では以下の用例が検出されている。

(a) 地盤が悪くては 意味 がないからです。
(b) ご主人に対してだけ対策してもあまり 意味 ないですよ。

一見、悪くない検出であるが、実は (a) は教師データの一つなので、(b) を検出するのは避けなければならない。教師付き LOF では、この問題を避けることができている。

ただし、SemEval-2 のデータではデータ数が少なく、しかも教師データとテストデータがほぼ同じ数あるという不自然な状況のために LOF において教師データの利用の効果が生じたとも考えられる。

(3) については SemEval-2 のデータに対する実験でも確認できた。本来、正常値の適切な生成モデルやクラスター形状が推定できれば、外れ値検出を精度良く行えると考えられるため、教師データからそれらを推定するアプローチは有効であると考えられる。

本手法の誤検出の原因について述べる。1 つは書き誤りに近いものである。例えば、以下は助詞が抜けていると見なすこともできる。

(c) 私が 子供 産んだ頃は、
(d) 忙しいでしょうから、お 時間 あるとき、

書き誤りは検出されてもしかたないし、この類の検出は有益性もあり問題は少ない。

他の誤検出の原因はいくつかあるが、複合語の認識の問題が大きい。名詞の語義識別の場合、対象単語が複合語の一部になっていれば、前後の単語の情報は語義識別の上での大きな情報となる。このため特異な複合語が検出されることが多い。検出された複合語が実際に専門性の高い用語である場合もあり、そのような場合には意味のある検出とも見なせるが、現在は複合語を単なる名詞連続で認識しているために以下のような検出が散見される。

(e) そんな 時間 必要ないけど、
(f) 給食費の未納がものすご〜く多い学校 現場 です。
(g) 加入 電話 サービスの基本料は

(e) は助詞が抜けて複合語と誤認識している。(f) や (g) などは専門用語か一般用語かの判断とも関わり、ここで行っているような単純な処理では解決は難しい。

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その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。これは今後の課題である。

もう一つ本手法の課題を述べておく。本手法では 3 つのパラメータが存在する。LOF における k-距離の k、LOF 値の上位いくつまでを候補に取るか、及び r_0 の値である。これらの値が異なると検出結果は全く異なってしまう。ここでの実験は予備実験を行い、適切そうな値を見積もって設定している。これらのパラメータはタスクに応じて、最適なものは異なるはずであり、これらパラメータの適切な設定方法が今後の課題である。

6 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手法である。人工的なデータや SemEval-2 の Japanese WSD タスクのデータを用いた実験により、提案手法の効果を示した。

提案手法には 3 つのパラメータが存在するので、それらを適切に設定する方法を考案することと、語義識別の精度を向上させる工夫を本研究に利用することが今後の課題である。

参考文献

[1] 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.

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

[3] 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.

[4] 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.

[5] 山西健司. データマイニングによる異常検知. 共立出版, 2009.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.