LOF と One Class SVM を用いた特異用例の検出

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からその単語の特異用例を検出する手法を提案する。

特異用例を厳密に定義することは不可能であるが、ここでは以下のような用例を特異用例と見なしている。

1. 用例中の対象単語の語義が新語義となっている
2. 用例中の対象単語を含む複合語が一般的な用語ではない

特異用例を検出することは、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。また特異用例はしばしば書き誤りとなっているので、誤りの検出としてもできる。

本論文では、特異用例を検出するために、特異用例を用例集中の外れ値と見なし、データマイニングにおける外れ値検出の手法 [6] を利用する。具体的には外れ値検出の代表的手法である Local Outlier Factor (LOF)[5] と One Class SVM[1] を組み合わせて利用する。

実験では、10 個の名詞を対象単語とし、BCCWJ コーパス [4] の「白書」からそれら対象単語の用例集を作成し、それら用例集に対して本手法を適用した。本手法は LOF や One Class SVM を単独で用いるよりも特異用例検出の正解率、再現率が優れていた。未検出と誤検出の主な原因は、用例間の類似度の測定が不適切なためである。この点の改善を今後の課題とする。

2　LOF と One Class SVM の組み合わせ

2.1　LOF

外れ値検出の手法は距離ベースの手法、密度ベースの手法、クラスタリングベースの手法等に分類できるが、LOF は密度ベースの代表的な手法である [5]。概略、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。

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

2.2　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。その結果、-1 のクラス側に属するデータを外れ値とする。

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

目的関数の式は以下である。

$$\min_{w,b,\xi,\rho} \frac{1}{2} w^T w - \rho + \frac{1}{\nu N} \sum_{i=1}^{N} \xi_i$$

subject to

$$w^T \phi(x_i) \geq \rho - \xi_i$$

$$\xi_i \geq 0 \quad (i = 1, 2, \cdots, N).$$

上記を以下の双対問題に変換して超平面を求める。

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu N}$$

$$\sum_{i=1}^{N} \alpha_i = 1.$$

2.3 LOF と One Class SVM の出力の積

LOF と One Class SVM は本タスク（用例集からの特異用例の検出）に関して、2つの問題がある。

第1の問題は両手法とも特異でない用例を数多く検出してしまう、すなわち、正解率が低いことである。本タスクにおいては、長さの短い用例が存在する、用例の総数が多くない、利用するコーパスが一般的でないこともある等が原因である。

第2の問題は、検出する用例の数を制御することが困難なことである。本タスクに対し、One Class SVM は用例集の1割から2割を特異用例として検出する。この数は明らかに多すぎる。また LOF ではある閾値を設定して、その閾値以上の $LOF(x)$ の値を持つ用例 x を特異用例とするが、この閾値を適切に設定することは難しい。

上記の2つの問題を解決するために、ここでは LOF と One Class SVM を組み合わせて利用する。具体的には、$LOF(x)$ の値から上位 20 個の用例を取り出し、それら用例と One class SVM から検出された用例との積を最終的な出力とする。つまり本手法では特異用例として検出する用例数は最大 20 である。この 20 という数値は本タスクでは十分に大きな値であることから設定した。

3 素性リスト

本タスクに LOF や One Class SVM を利用するには、まず用例を素性リストとして表現する必要がある。ここでは以下の4種類の素性を利用した。

e1: 直前の単語

e2: 直後の単語

e3: 対象単語の前後にあるそれぞれ 8 個までの名詞、ただし e1 と e2 のものは除く

e4: e1, e2 及び e3 に対するシソーラス番号

例を示す。対象単語を「核」として、以下の用例に対して作成される素性リストを示す。用例は形態素解析される。ここでは単語の境界には/（スラッシュ）を入れている。

日本/の/核/問題/を/議論/する/。

対象単語の「核」の直前の単語は「の」なので `"e1=の"` となり、直後の単語は「問題」なので `"e2=問題"` となる。「核」の前後の名詞をそれぞれ8個まで集めると前方からは `"e3=日本"` が取られ、後方からは `"e3=議論"` が取られる。後方からの `"e3=問題"` は `"e2=問題"` と重なるので取らない。

シソーラスとして、ここでは分類語彙表を用いる。分類語彙表で「日本」「議論」「問題」はそれぞれ 1.259_1、1.3070_3、1.3133_4 のシソーラス番号が与えられている。ここではコンマ以下が 3 桁のもの（1.259_1）からはコンマを除いた 4 桁の数（1259）を作り、コンマ以下が 4 桁のもの（1.3070_3、1.3133_4）からはコンマを除いた 4 桁の数（1307、1313）と 5 桁の数（13070、13133）を作った。以上より `"e4=1259"` 、 `"e4=1307"` 、 `"e4=1313"` 、`"e4=13070"` 、 `"e4=13133"` が作成される。

以上より、上記用例に対して以下の素性リストが作成できる。

```
{ e1=の, e2=問題, e3=日本, e3=議論,
  e4=1307, e4=13070, e4=1259, e4=1313,
  e4=13133 }
```

4 実験

ここではコーパスとして BCCWJ コーパス [4] の「白書」を利用する。対象単語としては、以下に示す10個の名詞を用いた。これらの名詞は Senseval2 日本語辞書タスク [2] で用いられた名詞の一部である。

核、 一般、記録、時間、市民、 時代、情報、精神、代表、民間

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

本手法の再現率を評価するために、各対象単語に対して 1 つ人工的に特異用例を作成し、それらをコーパスに含めた。各対象単語に対して作成した特異用例は以下の通りである。

核： 再生核ヒルベルト空間の概念を理解する。
一般：連続関数と一般変換群の関係。
記録： 近年、年金記録問題が騒がれている。
時間： 過去と未来がつながる円環時間。
市民： 市民大学講座で統計学を学ぶ。
時代： 「雑居時代」は昔のホームドラマです。
情報： 形態情報端末が普及した。(誤り)
精神： オリンピック精神で世界が感動。
代表： これは熊本産の代表メロンです。
民間： 明治の民間数学者松岡文太郎の仕事と功績。

対象単語「核」に対する検出結果を示す。コーパスから得られた「核」に対する用例数は 1,031 であった。この用例集から、本手法は以下の 8 つの用例を特異用例として検出した。

(1) 再生核ヒルベルト空間の概念を理解する。(○)
(2) 核テナントに，必要に応じ … 出店させ，… (○)
(3) 日米間で移転される核質物に対するものとして … (○)
(4) …，核移植，遺伝子組換え等の研究開発を … (×)
(5) 西ドイツの場合も核エネルギーが主体であるが，… (×)
(6) 昭和五十七年核廃棄物政策法が成立し，… (×)
(7) 高レベル放射性廃棄物に含まれる核種を分離し、… (○)
(8) 業務核都市に対する支援措置として，… (○)

上記検出例における ○ と × は、それぞれ正しい検出と判定したもの、誤った検出と判定したものを示している。特異用例と判定したものは用例 (1)、(2)、(3)、(7)、(8) の 5 つである。用例 (1) は人工的に加えた特異用例である。用例 (3) の「核質物」は「核物質」の誤りである。用例 (8) の「核都市」は「核都市」の古い表記である。用例 (2) の「核テナント」や用例 (7) の「核種」は専門用語であり、一般的な文書には現れない。その他の用例 (4)、(5)、(6) は特異用例でないと判定した。ここから正解率は $5/8 = 0.625$ となる。

その他の単語の結果もまとめたものを表 1 に示す。

表 1: 正解率

単語	用例数	本手法	LOF	OC-SVM
核	1,031	5 (8)	10 (20)	4×5.25 (105)
一般	2,047	1 (8)	3 (20)	2×9.85 (197)
記録	326	2 (4)	4 (20)	3×2.25 (45)
時間	1,411	1 (4)	3 (20)	1×7.90 (158)
市民	210	2 (9)	3 (20)	2×2.85 (57)
時代	289	3 (8)	7 (20)	7×2.65 (53)
情報	3,678	2 (2)	6 (20)	2×9.25 (185)
精神	432	0 (5)	4 (20)	0×3.60 (72)
代表	351	3 (8)	7 (20)	2×3.50 (70)
民間	1,474	2 (7)	2 (20)	2×7.60 (152)
正解率		21/63 = 0.333	49/200 = 0.245	120.3/1094 = 0.110

表 1 は本手法、LOF、OC-SVM (One Class SVM) の正解率を示している。各単語に対して、それぞれの手法が検出した用例数を括弧内に示し、その中で実際に特異用例と判断できるものの数をその左に示した。ただし OC-SVM に関しては以下の形で表記している。

$$a \times b\ (\ s\ )$$

s は検出した用例数である。検出した用例すべてに対して実際にそれが特異用例かどうかを判断するのは用例数が多いために困難である。そのために検出した用例からランダムに 20 用例を取り出し、それらに対して実際にそれが特異用例かどうかを判断した。a はその 20 用例中で特異用例と判断された用例数である。b は $s/20$ を意味している。

表 1 から本手法が LOF や One Class SVM よりも正解率が高いことがわかる、

次に人工的に作成した特異用例が各手法で検出できたかどうかを調べることで擬似的に再現率を評価する。この結果を表 2 に示す。

表 2: 再現率

単語	本手法	LOF	OC- SVM
核	○	1	○
一般	×	1,135	○
記録	×	32	○
時間	×	21	×
市民	×	105	○
時代	×	3	×
情報	×	3,379	×
精神	×	43	○
代表	×	39	×
民間	×	117	○
再現率	1/63 = 0.016	2/200 = 0.010	6/1094 = 0.005

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

表 2 の ○ と × は各単語に対して人工的に作成した特異用例が検出できたかどうかを示している。○ は検出できたことを意味し、× はできなかったことを意味する。LOF の場合、数値が記されているが、これは人工的に作成した特異用例の $LOF(x)$ 値の順位である。つまりこの数値が 20 以下であれば検出できたことを意味し、そうでなければ検出できなかったことを意味する。各手法について、○の数を検出した用例数で割ることで、擬似的に再現率を求めた。この擬似的な再現率に関しても本手法は LOF や One Class SVM よりも優れていた。

5 考察

本手法の未検出や誤検出の主な原因は用例間の類似度を適切に設定できていないからである。本手法は教師なし学習の一種であるため、用例間の類似度の測定法はアドホックにならざるおえない。アドホックでなく類似度を適切に測るためには、（半）教師あり学習が必要である [7]。

また本タスクに関しては、用例間の類似度を単一の測定法で測ることも困難である。本論文では、以下の２つのタイプを特異用例とみなしている。

1. 用例中の対象単語の語義が新語義となっている
2. 用例中の対象単語を含む複合語が一般的な用語ではない

一般にタイプ (2) の特異用例はタイプ (1) の特異用例ではない。本手法ではタイプ (2) の用例を検出する傾向があり、そのため、タイプ (1) の特異用例は検出できない。実際に、本実験においてもタイプ (1) の特異用例は検出できていない。両者の特異用例を検出ためには、各々のタイプに対する類似度を設定する必要がある。

また本手法を改善するためには、One Class SVM を効果的に利用すべきである。One Class SVM はデータのベクトル表現方法やカーネル関数の選択に敏感である [3]。実際にここでの実験では One Class SVM は、それほど有効に機能していなかった。One Class SVM を効果的に利用するために、今後は用例のベクトル表現方法を改善する予定である。

最後に、本タスクは利用するコーパスを評価できる可能性があることを指摘しておく。自然言語処理システムの研究開発のために、これまで多くのコーパスが構築されてきたが、構築したコーパスを評価することは困難である。一方、バランスのとれたコーパスには特異用例が存在しないと考えられるので、特異用例が存在するかどうかでそのコーパスがバランスが取れているかどうかを評価できる。

6 結論

本論文ではコーパスから対象単語の特異用例を検出するために特異用例を用例集中の外れ値とみなし、データマイニングの外れ値検出の手法である LOF と One Class SVM を組み合わせて利用した。実験では 10 個の名詞を対象単語として特異用例の検出を行った。本手法は LOF や One Class SVM を単独で利用するよりもよいパフォーマンスを示した。今後は用例間の類似度の測定法を改良したい。また用例のベクトル表現方法を改良し、One Class SVM を有効に利用したい。

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] Kiyoaki Shirai. SENSEVAL-2 Japanese Dictionary Task. In *SENSEVAL-2*, pp. 33–36, 2001.

[3] Larry M. Manevitz and Malik Yousef. One-class svms for document classification. *Journal of Machine Learning Research*, Vol. 2, pp. 139–154, 2002.

[4] Kikuo Maekawa. Design of a Balanced Corpus of Contemporary Written Japanese. In *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58, 2007.

[5] 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.

[6] Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, Vol. 22, pp. 85–126, 2004.

[7] Liu Yang. An Overview of Distance Metric Learning. In *Technical report, Michigan State University*, 2007.

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.