Detection of Peculiar Examples using LOF and One Class SVM

Hiroyuki Shinnou, Minoru Sasaki

Ibaraki University, Department of Information and Computer Science
4-12-1 Nakanarusawa, Hitachi, Ibaraki, Japan 316-8511
{shinnou, msasaki}@mx.ibaraki.ac.jp

Abstract

This paper proposes the method to detect peculiar examples of the target word from a corpus. The peculiar example is regarded as an outlier in the given example set. Therefore we can apply many methods proposed in the data mining domain to our task. In this paper, we propose the method to combine the density based method, Local Outlier Factor (LOF), and One Class SVM, which are representative outlier detection methods in the data mining domain. In the experiment, we use the Whitepaper text in BCCWJ as the corpus, and 10 noun words as target words. Our method improved precision and recall of LOF and One Class SVM. And we show that our method can detect new meanings by using the noun 'midori (green)'. The main reason of un-detections and wrong detection is that similarity measure of two examples is inadequacy. In future, we must improve it.

1. Introduction

This paper proposes the method to detect peculiar examples of the target word from a corpus.
It is impossible to define a peculiar example strictly. However, in this paper we regard following examples as peculiar examples:

(1) a meaning of the target word in the example is new,

(2) a compound word consisting of the target word in the example is new or very technical.

The detection of peculiar examples is useful to construct a dictionary and training data for the word sense disambiguation (WSD) task. In addition, some clerical errors are detected as peculiar examples, so this detection system can be used as an error detection system.
The peculiar example is regarded as an outlier in the given example set. Therefore we can apply many methods proposed in the data mining domain to our task(Victoria J. Hodge and Jim Austin, 2004). In this paper, we propose the method to combine the density based method, Local Outlier Factor (LOF)(Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander, 2000), and One Class SVM(B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson, 2001), which are representative outlier detection methods in the data mining domain.
In the experiment, we use the Whitepaper text in BCCWJ(Maekawa, 2007) as the corpus, and 10 noun words as target words. Our method improved precision and recall of LOF and One Class SVM. And we show that our method can detect new meanings by using the noun '緑 (green)'. The main reason of un-detections and wrong detection is that similarity measure of two examples is inadequacy. In future, we must improve it.

2. Combination of LOF and One Class SVM

2.1. LOF

LOF is a density based outlier detection method(Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander, 2000).
It is trivial that a far object from all other objects, like the object *A* in the figure 1, is outlier. The distance based method uses this characteristic to detect outliers. However, the distance based method is not enough, because it cannot detect an outlier like the object *B* in the figure 1. The object *B* is not so far from the near cluster, but the density of that cluster is very high. The density based method detects outliers like the object *B*.

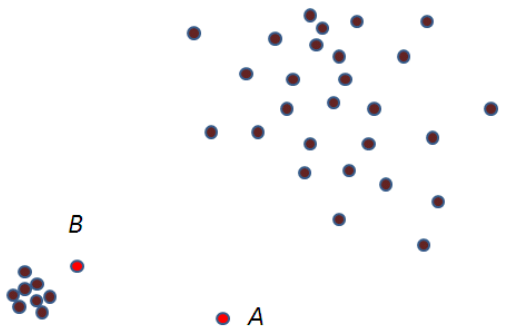

Figure 1: density based method and distance based method

In LOF, the outlier degree of the object is measured using the density of neighborhood of an object. We can detect outliers from that degree.
In order to define LOF, first we must define a distance called $kdist(x)$ for an object x. The $kdist(x)$ is defined as the distance $d(x, o)$ between x and an object $o \in D$ such that:

1. for at least k objects $o' \in D \setminus \{x\}$ it holds that $d(x, o') \leq d(x, o)$, and
2. for at most $k - 1$ objects $o' \in D \setminus \{x\}$ it holds that $d(x, o') < d(x, o)$.

Generally the object o means the k-th nearest object from the object x. Above definition can cope with the case that some objects are equal to the object x.
By the $kdist(x)$, $N_k(x)$, $rd_k(x,y)$ and $lrd_k(x)$ are defined as follows:

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

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

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

By using them, LOF is defined as below.

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

2.2. One Class SVM

One Class SVM is an outlier detection method using $\nu-$SVM(B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson, 2001). Classes of all objects are set to +1, and the class of the origin is set to -1. Under this setting, $\nu-$SVM provides the hyperplane dividing two classes, and -1 side objects are regarded as outliers.
The figure 2 viscerally explains this mechanism. The more the hyperplane gets close to the origin, the higher the precision of SVM is. The more the hyperplane gets away to the origin, the bigger the margin between the hyperplane and support vectors is. Thus, the optimal position of the hyperplane is calculated.

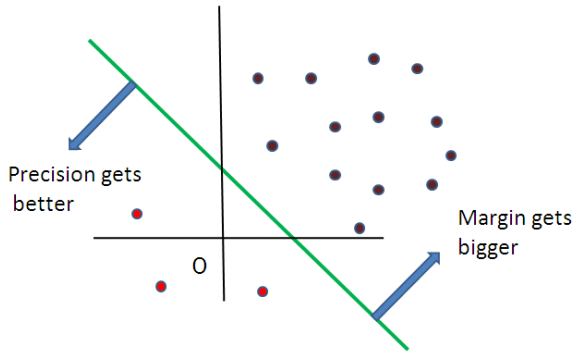

Figure 2: One Class SVM

The primal form is following:

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

subject to

$$w^T \phi(x_i) \geq \rho - \xi_i$$
$$\xi_i \geq 0 \ \ (i = 1, 2, \cdots, N).$$

The dual is:

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

subject to

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

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

2.3. Combination of LOF and One Class SVM

LOF and One Class SVM have two problems for our task. One problem is that both methods detect many non-outliers objects, that is, precisions are low. In our task, some examples are short sentences, the number of examples is small, and the used corpus is not always general. LOF and One Class SVM are affected by these problems.
Another problem is that it is difficult to control the number of detections. In the case of our task, One Class SVM detects 10% to 20% of all examples as outliers. LOF judges objects with the higher LOF value than a threshold as outliers. It is difficult to set up the proper threshold .
To overcome above two problems, we propose the method to combine LOF and One Class SVM. Specifically, first we pick up highest 20 objects of LOF values, then we output the union of these objects and outputs from One class SVM. So the maximum number of detections by our method is 20. This number is enough large for our task.

3. Features

To perform our method, we must convert the object to the feature list. In this paper, we use following 4 types of features.
Suppose the target word is $w = w_i$, which is the i-th word in the sentence.

e1: the word w_{i-1}
e2: the word w_{i+1}
e3: each 8 noun words in front of and behind w_i
e4: thesaurus ID number of e1, e2 and e3

For example, let's consider the following sentence[1] in which the target word is w_3 ='核 *(kaku)*' [2].

日本/の/核/問題/を/議論/する/.
nihon/no/kaku/mondai/wo/giron/suru/.

The previous word of the target word is w_2 ='の *(no)*', so `e1=no`. The behind word of the target word is w_4 ='問題 *(mondai)*', so `e2=mondai`. As noun words in front of and behind of the target word, w_1 ='日本 *(nihon)*' and w_6 ='議論 *(giron)*' are picked up, so `e3=nihon` and `e3=giron`.

[1] A sentence is segmented into words, and each word is transformed into its original form by morphological analysis.
[2] The word '核 (kaku)' has at least two meanings: "center" and "nuclear".

Next we look up the thesaurus ID of the word *'mondai'*, and find `1.3070_3` [3]. In our thesaurus, a higher number corresponds to a higher level meaning.
In this paper, we use a four-digit number and a five-digit number of a thesaurus ID. As a result, for `'e2=mondai'`, we get `'e4=1307'` and `'e4=13070'`. In the same way, for `'e3=giron'` and `'e3=nihon'`, we get `'e4=1313'`, `'e4=13133'` and `'e4=1259'`. Note that we do not look up the thesaurus ID for `'e1=no'` because the word *'no'* is not a noun word.
As a result, the system generates the following feature list from the above example.

```
{ e1=no, e2=mondai, e3=nihon, e3=giron,
  e4=1307, e4=13070, e4=1259, e4=1313,
  e4=13133 }
```

4. Experiments

As a corpus in the experiment, we used the Whitepaper text, which is a part of the Japanese BCCWJ corpus(Maekawa, 2007). As the target word, we selected following 10 noun words, which are a part of words used in the Senseval2 Japanese dictionary task(Kiyoaki Shirai, 2001).

核 (kaku), 一般 (ippan), 記録 (kiroku),
時間 (jikan), 市民 (shimin), 時代 (jidai),
情報 (joho), 精神 (seishin), 代表 (daihyo),
民間 (minakn)

In order to evaluate the recall of our method, we made the artifical outlier example for each word, and added it to the corpus.
The added outlier examples are as follows:

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

Following is the result for the target word '核 (kaku).'

(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) 細胞融合，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) 業務核蔀市に対する支援措置として，… (○)

The number of examples of the target word '核 (kaku)' is 1,031. From these examples, we detected above 8 examples. The sign ○ and × means that the example is peculiar or not respectively. This judgment is conducted subjectively. However, we believe that this judgment can be approved. The (1) is the artifical outlier example. The '核質物' in (3) is the clerical error of '核物質.' The '核蔀市' in (8) is the old form of '核都市.' The '核テナント' in (2) and the '核種' in (7) are seldom used in a general document.

The Table 1 shows the whole result of the experiment.

Table 1: Precision

word	examples	Our Method	LOF	OC-SVM
kaku	1,031	5 (8)	10 (20)	4×5.25 (105)
ippan	2,047	1 (8)	3 (20)	2×9.85 (197)
kiroku	326	2 (4)	4 (20)	3×2.25 (45)
jikan	1,411	1 (4)	3 (20)	1×7.90 (158)
shimin	210	2 (9)	3 (20)	2×2.85 (57)
jidai	289	3 (8)	7 (20)	7×2.65 (53)
joho	3,678	2 (2)	6 (20)	2×9.25 (185)
seishin	432	0 (5)	4 (20)	0×3.60 (72)
daihyou	351	3 (8)	7 (20)	2×3.50 (70)
minkan	1,474	2 (7)	2 (20)	2×7.60 (152)
precision		21/63 = 0.333	49/200 = 0.245	120.3/1094 = 0.110

In the Table 1, 'examples' means the number of examples extracted from the corpus for each word. 'Our Method' and 'LOF' mean the number of correct detections by our methods and LOF respectively. The number in parenthesis means the number of whole detections in each method. 'OC-SVM' means the result of detection by One Class SVM. This is expressed the following form:

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

The s means the number of whole detections. It is hard to check to be peculiar or not for all these examples. Therefore we pick up 20 examples at random from detections for each word, and we conduct the check for only them. The a means the number of peculiar examples in 20 examples, and $b = s/20$. The Table 1 shows that our method improves precision of LOF and One Class SVM.
Then we investigated the artifical peculiar examples to be detected or not. The result is shown in Table 2.

Table 2: Recall

word	Our Method	LOF	OC- SVM
kaku	○	1	○
ippan	×	1,135	○
kiroku	×	32	○
jikan	×	21	×
shimin	×	105	○
jidai	×	3	×
joho	×	3,379	×
seishin	×	43	○
daihyou	×	39	×
minkan	×	117	○
recall	1/63 = 0.016	2/200 = 0.010	6/1094 = 0.005

[3] In this paper we use the *bunrui-goi-hyou* as a Japanese thesaurus.

The sign ○ and × means that the artifical peculiar example was detected and not respectively. The number in the column of 'LOF' is the rank of the LOF values. That is, if the number is less than 20, it means that LOF detected the artifical peculiar example. We approximately define the recall to be the ratio of the number of ○ for the number of whole detections. Also from the view of our recall, our method is better than LOF and One Class SVM.

In above experiments, new meanings of target word are not detected. This is because our above 10 words are too general, so new meanings of these words do not exists in the corpus. Actually, new meaning are too rare in any corpus, so it is difficult to evaluate a method to detect new meanings.
Fortunately, we have data with new meaning tags, which will be used in the Japanese WSD task in SemEval-2[4]. For example, the meaning 'plant' of the noun '緑 (midori)' [5] is taken as new in that data.
Here, we set the noun '緑' as the target word, and try our method. The used corpus is same to the above experiments. As a result, the number of example sentences is 387, and our method detected following 11 examples as peculiar examples.

(1) ‥，緑黄野菜等他の食品 …（○）
(2) ‥，緑が少なく水や空気が汚れている（○）
(3) ‥ 下水道水緑景観モデル事業、…　（○）
(4) ‥「緑サポーター」養成研修、…　（○）
(5) ‥ 、横浜市港北区、緑区などが ‥　（○）
(6) ‥「緑住まちづくり推進事業」を推進した ‥　（○）
(7) ‥ 栃木県知事は，有限会社池田緑商店から ‥　（○）
(8) ‥ 栃木県知事は，有限会社池田緑商店に対し，‥　（○）
(9) ‥ 都市における緑は、気温の調節、…　（○）
(10) ‥ 森林や緑に対する国民の関心を …　（○）
(11) ‥ 申請人有限会社池田緑商店は，…　（○）

The '緑黄野菜' in (1) is mistype of '緑黄色野菜.' Compounds including the word '緑' in (3), (4) and (6) are not general. The word '緑' in (5) is a location name, and the word '緑' in (7), (8) and (11) are person names. Meanings of the word '緑' in (2), (9) and (10) are 'plant', so new.
From this experiment, we can confirm that our method can detect new meanings.

5. Discussions

The main reason of error detections and un-detections of our method is that the similarity measure of examples is not precise. Our method is a kind of unsupervised learning, so the similarity measure cannot avoid to be ad-hoc. To measure the similarity precisely, we believe that (semi-)supervised learning is needed(Yang, 2007).
Moreover, in this paper, the peculiar example is defined by following two types:
(1) the meaning of the target word in the example is new,
(2) a compound word consisting of the target word in the example is new or very technical.
However, to detect these examples, it is difficult to use the uniform similarity measure. Generally, an example of type 2 is not type 1. Our method tends to detect examples of type 2. Actually our method detected type 1 outliers in additional experiment. However the detected new meaning 'plant' is not new actually. The Japanese WSD task in SemEval-2 uses the Iwanami Kokugo Jiten, a Japanese dictionary published by Iwanami Shoten, where this meaning meaning of the noun '緑' does not exists. Other dictionary has that meaning of the noun '緑.' In future, we need define the similarity measure for each type.
To improve our method, we use One Class SVM effectively. One Class SVM is sensitive to the choice of representation and kernel(Larry M. Manevitz and Malik Yousef, 2002). Actually One Class SVM did not work so well in the experiment. To use One Class SVM well, we need improve the representation of example.
At last, we note that our task can evaluate a corpus. For natural language processing systems, many corpora have been constructed. However, it is difficult to evaluate the corpus. On the other hand, the balanced corpus can judge the peculiar example. Thus, a system detects a peculiar example, and the example is judged to be peculiar by the balanced corpus. If that judgment is equivalent to human judgment, it means that the corpus is balanced.

6. Conclusion

This paper proposed the method to detect peculiar examples by combining LOF and One Class SVM. In the experiment using 10 target words, our method provided the better score in both of precision and recall than LOF and One Class SVM. In future, we need improve the similarity measure and use One Class SVM well by improving representation of an example.

7. References

B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.

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

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

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

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

Victoria J. Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126.

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

[4]The task number is 16. Refer to http://semeval2.fbk.eu/.

[5]The Japanese word '緑' corresponds to the 'green' in English.