Semi-supervised learning by Fuzzy clustering and Ensemble learning

Hiroyuki Shinnou, Minoru Sasaki

Dept. of Systems Engineering, Ibaraki University,
4-12-1 Nakanarusawa, Hitachi,
Ibaraki JAPAN 316-8511
shinnou@dse.ibaraki.ac.jp

Dept. of Computer and Information Sciences,
Ibaraki University, 4-12-1 Nakanarusawa, Hitachi,
Ibaraki JAPAN 316-8511
sasaki@cis.ibaraki.ac.jp

Abstract

This paper proposes a semi-supervised learning method using Fuzzy clustering to solve word sense disambiguation problems. Furthermore, we reduce side effects of semi-supervised learning by ensemble learning. We set N classes for N labeled instances. The n-th labeled instance is used as the prototype of the n-th class. By using Fuzzy clustering for unlabeled instances, prototypes are moved to more suitable positions. We can classify a test instance by the k Nearest Neighbor (k-NN) with the moved prototypes. Moreover, to reduce side effects of semi-supervised learning, we use the ensemble learning combined the k-NN with initial labeled instances, which is initial prototype, and the k-NN with prototypes moved by Fuzzy clustering.

1. Introduction

In this paper, we propose a semi-supervised learning method using Fuzzy clustering to solve word sense disambiguation problems. Furthermore, we reduce side effects of semi-supervised learning by ensemble learning.

Many problems in natural language processing can be converted into classification problems, and be solved by an inductive learning method. This strategy has been very successful, but it has a serious problem in that an inductive learning method requires labeled data, which is expensive because it must be made manually. To overcome this problem, semi-supervised learning methods using huge unlabeled data to boost the performance of rules learned by small labeled data have been proposed recently(Nigam et al., 2000)(Blum and Mitchell, 1998)(Yarowsky, 1995).

In this study, we propose a semi-supervised learning method using Fuzzy clustering. Fuzzy clustering allows a instance to belong multiple classes with the degree of membership for the class. First, we set N classes for N labeled instances. Next, we regard the n-th labeled instance as the prototype of n-th class. Therefore, each initial cluster has only one element, that is the labeled instance. Next, by using the Fuzzy clustering method, we compute the degree of membership for the class of an unlabeled instance. As the result, prototypes move to more suitable points. In actual classification, we use k Nearest Neighbor method (k-NN) with the prototypes moved by Fuzzy clustering. These prototypes can be regarded as new labeled data. This method is a kind of semi-supervised methods because our new labeled data is constructed by using unlabeled data.

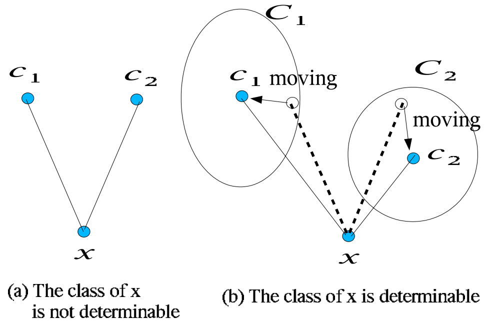
However, semi-supervised learning does not always improve the rules learned through the provided labeled data, and often degrades the rules. We refer to this problem as the *side effect* of semi-supervised learning in this paper. To reduce the side effect, we use ensemble learning. Our ensemble learning combines the k-NN with initial labeled data and the k-NN with new labeled data obtained by Fuzzy clustering. Ensemble learning makes up for weakness of each learning method, so it recovers side effects of semi-supervised learning.

In experiments, we took verb words of the Japanese Dictionary Task in SENSEVAL2(Kurohashi and Shirai, 2001). The standard k-NN, that is a supervised learning method, achieved the accuracy 77.79%, and semi-supervised learning by Fuzzy clustering achieved the accuracy 77.83%. These values are not so different because some side effects were produced. By ensemble learning, we improved the accuracy from 77.83% to 78.53%.

2. Solution of WSD by Fuzzy clustering

2.1. Semi-Supervised learning by Fuzzy clustering

First we give an intuitive explanation why clustering can boost the performance of the rules. Let's consider the situation shown in the figure 1.



(a) The class of x is not determinable

(b) The class of x is determinable

Figure 1: Movement of prototypes by clustering

Let $dis(\cdot, \cdot)$ be a distance measure function between two vectors. In figure 1 (a), x is a test instance, c_1 is the prototype of the class C_1, and c_2 is the prototype of the class C_2. We cannot judge whether the x belongs to C_1 or C_2, because $dis(x, c_1) = dis(x, c_2)$. On the other hand, in figure 1 (b), c_1 and c_2 move to more suitable points by clustering. As the result, we can judge the class of x.

However, it is not so simple because a class of an unlabled instance is vague. Therefore, we use Fuzzy Clustering to cope with the vagueness.

Next we explain an algorithm of Fuzzy clustering(Miyamoto, 1999). Let

$$C = \{C_1, C_2, \cdots, C_m\}$$

be a set of classes. This means that we have m labeled instances initially. Suppose we have n instances, for which we do clustering. This means that the total number of all labeled data and all unlabeled data is n[1]. The instance is expressed by a point on p dimension Euclid space, and the k-th instance is expressed by the following column vector.

$$x_k = (x_k^1, \ldots, x_k^p)^T$$

We define u_{ik} the degree of membership for the class C_i of the instance x_k. In standard clustering, u_{ik} is 0 or 1, but in Fuzzy clustering, it is a real number in $[0, 1]$. We define the $m \times n$ matrix U whose (i, k) element is u_{ik}. Let

$$v_i = (v_i^1, \ldots, v_i^p)^T.$$

be the prototype of the class C_i. We define the $p \times m$ matrix V whose (i, k) element is v_i^k.

Fuzzy clustering updates V and U step by step to minimum a target function $J(U, V)$. In Fuzzy clustering, it is a problem what target function we should use. In this paper, we use the following standard target function:

$$J(U, V) \quad = \quad \sum_{k=1}^{n} \sum_{i=1}^{m} (u_{ik})^r \|x_k - v_i\|^2 \qquad (1)$$

where $r > 1$. We set the r to be 2.

In the figure 2, we show the algorithm **FCM** to update V and U. This algorithm is called as *Fuzzy c-means*.

```
FCM (Fuzzy c-means)

step 1. Set the initial $\bar{V}$.

step 2. Fix $\bar{V}$, and solve

            $\min_{U \in M_f} J(U, \bar{V})$.

        Set that solution to $\bar{U}$.

step 3. Fix $\bar{U}$, and solve

            $\min_{V} J(\bar{U}, V)$.

        Set that solution to $\bar{V}$.

step 4. If $(\bar{U}, \bar{V})$ is convergent, this algorithm finishes. If no so, go back to step 2.
```

Figure 2: Fuzzy c-means

The solution in the step 2 can be obtained as follows. If x_k is not equal to all v_i,

$$u_{ik} \quad = \quad \left[\sum_{j=1}^{m} \left(\frac{\|x_k - \bar{v}_i\|^2}{\|x_k \; - \bar{v}_j\|^2} \right)^{\frac{1}{r-1}} \right]^{-1} \quad , \; for \; x_k \neq v_i, i = 1, \ldots, m \qquad (2)$$

If x_k is equal to a certain v_i,

$$u_{ik} = 1; \quad u_{jk} = 0 \; (j \neq i). \qquad (3)$$

The solution in the step 3 can be obtained as follows:

$$v_i = \frac{\sum_{k=1}^{n} (\bar{u}_{ik})^r x_k}{\sum_{k=1}^{n} (\bar{u}_{ik})^r}. \qquad (4)$$

The convergence condition of the step 4 is also various. In this paper, we take the simplest way which the maximum iteration number is set. We set the maximum iteration number to be 5.

Note how to handle labeled instances in the **FCM**. In the first loop, a labeled instance is equal to a prototype. Thus, we use the equation 3 in the step 2. After the first loop, a labeled instance is not generally equal to all prototypes. However, we use not the equation 2 but the equation 3 even in this case because the class of the labeled instance is fixed.

2.2. Features

We need to express an instance as a p dimensional vector. In general, an instance in a classification problem is expressed by the feature list. Therefore, we express an instance as the p dimensional vector by regarding each feature as each dimension. If the instance has the k-th feature, the value in the k-th dimension is set to be 1. Conversely, if the instance does not have the k-th feature, the value in the k-th dimension is set to be 0.

To make a feature list for WSD, we use following six attributes (`e1` to `e6`) in this paper. Suppose that the target word is w_i which is the i-th word in the sentence.

e1: the word w_{i-1}
e2: the word w_{i+1}
e3: two content words in front of w_i
e4: two content words behind w_i
e5: thesaurus ID number of e3
e6: thesaurus ID number of e4

For example, we make features from the following sentence [2] in which the target word is *'kiroku'*[3].

kako/saikou/wo/kiroku/suru/ta/.

Because the word to the left of the word *'kiroku'* is *'wo'*, we get `'e1=wo'`. In the same way, we get `'e2=suru'`. Content words to the left of the word *'kiroku'* are the word *'kako'* and the word *'saikou'*. We select two words from them in the order of proximity to the target word. Thus, we get `'e3=kako'` and `'e3=saikou'`. In the same way, we get `'e4=suru'` and `'e4=.'`. Note that the comma and the period are defined as a kind of content words in this paper. Next we look up the thesaurus ID of the word *'saikou'*, and find `3.1920_4` [4]. In our thesaurus, as shown in Figure 3, a higher number corresponds to a higher level meaning.

[1] A class of a labeled instance is fixed, but Fuzzy clustering needs labeled instances to compute the degree of membership for the class for an unlabeled instance.

[2] A sentence is segmented into words, and each word is transformed to its original form by morphological analysis.

[3] *'kiroku'* has at least two meanings: 'memo' and 'record'.

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

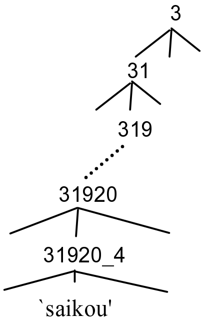


Figure 3: Japanese thesaurus: Bunrui-goi-hyou

In this paper, we use a four-digit number and a five-digit number of a thesaurus ID. As a result, for `'e3=saikou'`, we get `'e5=3192'` and `'e5=31920'`. In the same way, for `'e3=kako'`, we get `'e5=1164'` and `'e5=11642'`. Following this procedure, we should look up the thesaurus ID of the word '*suru*'. However, we do not look up the thesaurus ID for a word that consists of *hiragana* characters, because such words are too ambiguous, that is, they have too many thesaurus IDs. When a word has multiple thesaurus IDs, we create a feature for each ID.

As a result, we get following ten features from the above example sentence:

```
e1=wo, e2=suru, e3=saikou, e3=kako,
e4=suru, e4=.,  e5=3192, e5=31920,
e5=1164, e5=11642.
```

3. Side effects and ensemble learning

In general, semi-supervised learning suffers the side effect problem.

We use ensemble learning to reduce side effects of semi-supervised learning. In ensemble learning, multiple classifiers are obtained through multiple learning methods. Each classifier judges a class of a test instance. The final judgment for the test instance is done by considering all the various judgments together. Ensemble learning makes up for weakness of each learning method, so the accuracy of ensemble learning tends to be better than the accuracy of each learning method(Ueda and Nakano, 1997).

In this paper, we combine the supervised learning method and the semi-supervised learning method. First, we learn the classifier A through the initial labeled data and the classifier B through the new labeled data obtained by Fuzzy clustering. In our method, the classifier A corresponds to the k-NN using the initial labeled data, and the classifier B corresponds to the k-NN using the new labeled data obtained by Fuzzy clustering.

The question is how to combine two classifiers. In this paper, we do it by a weighted majority vote of $2k$ answers from the classifier A and B. To put it concretely, we pick up each k classes with weight by the classifier A and B. Finally, we add together the weight in each class, and we output the class with the biggest weight. Note that we use the dot product as the weight.

4. Experiments

To confirm the effectiveness of our methods, we tested with 50 verbs of the Japanese Dictionary Task in SENSEVAL2(Kurohashi and Shirai, 2001).

The Japanese Dictionary Task is a set of standard WSD problems. As the evaluation words, 50 noun words and 50 verb words are provided. We use only 50 verb words. The number of labeled instances and unlabeled instances is 172.7 and 6571.9 on average respectively.

Table 1 shows the result of experiments. In this table, the column of **k-NN** and the column of **Fuzzy** means the mean of 50 accuracy (%) of the k-NN using initial labeled instances and the mean of 50 accuracy (%) of the k-NN using new labeled instances obtained by Fuzzy clustering, respectively. In both cases, k was fixed to 5. The column of **Ensemble** means the mean of 50 accuracy (%) of the proposed ensemble method, that is the ensemble of above two methods (**k-NN** and **Fuzzy**).

Table 1: Result of experiments (accuracy (%))

k-NN	Fuzzy	Ensemble
77.79%	77.83%	78.53%

This table shows the effectiveness of our method.

5. Discussion

5.1. Effectiveness of ensemble learning

One way to overcome the side effect of semi-supervised is cross validation(Shinnou and Sasaki, 2003). By using cross validation, we can estimate whether semi-supervised learning is valid for the focused problem or not.

In this paper, we proposed another method to reduce side effects, that is ensemble learning of supervised learning and semi-supervised learning. In the experiment, the accuracy of the word '*motomeru*', '*tyukau*' and '*ukeru*' fell 8.0%, 9.75% and 5.5% by the semi-supervised learning respectively. However, ensemble learning brought back the base line, and curbed the big loss of the accuracy.

There were 22 words in all evaluation words (50 words), for which the ensemble learning improved the semi-supervised learning. For 18 words, accuracy of the ensemble learning and semi-supervised learning are equal. For remained 10 words, the ensemble learning degraded the semi-supervised learning. This shows that the ensemble learning reduces side effects of semi-supervised learning effectively. Moreover, there were just only 5 words in the degraded 10 words, for which the accuracy of the ensemble learning is worse than the accuracy of the supervised learning. Therefore, the side effect of ensemble learning is small.

5.2. Use of other based supervised method

Semi-supervised learning using Fuzzy clustering has the advantage that new labeled data is generated. We can use various supervised learning methods by using this new labeled data. However, methods except for k-NN may not be available. In fact, we conducted the experiments using decision list method and Naive Bayes method. In the case of the decision list, a slight improvement was achieved, but there is no improvement in the case of Naive Bayes. We think that the cause is the incompatibility of the distance measure. Fuzzy clustering uses a distance measure. That measure is incompatible with decision list method and Naive Bayes. On the other hand, we uses k-NN with the same distance measure as the Fuzzy clustering. Therefore our method succeeded. If we use another learning method besides k-NN, we have to use the suitable distance measure in the clustering stage. We can regard the method combining Naive Bayes and EM algorithm as a kind of clustering methods (Nigam et al., 2000).

5.3. Related works

Co-training(Blum and Mitchell, 1998) is a powerful semi-supervised learning method. Co-training requires two independent feature sets. First it constructs a classifier through one feature set. The classifier assigns classes to instances in an unlabeled data set, and then some instances with reliable labels are picked up. These instances are added to the labeled data set. By the same procedure, another feature set is used to add some instances to the labeled data. By iterating these procedures, Co-training augments the labeled data, thereby improving the accuracy of the learned classifier.

However, Co-training has some serious problems. The biggest problem is that it is difficult to set up two independent feature sets. Furthermore, Co-training requires consistency besides independence for two feature sets. This condition makes it difficult to apply Co-training to multiclass classification problems. On the other hand, our method can be applied to multiclass classification problems without any modification. Therefore, our method is more practical than Co-training.

Yarowsky proposed the semi-supervised learning method for WSD(Yarowsky, 1995). His method is reported to be a special case of Co-training(Blum and Mitchell, 1998). As two independent feature sets, one is the context surrounding the target word and the other is the heuristic of ‘one sense per discourse’. However, it is unknown how valid this heuristic is for granularity of meanings of our evaluation words. Furthermore, this method needs documents in which the target word appears multiple times, as unlabeled data. Therefore, it is not so easy to gather unlabeled data. On the other hand, our method does not have such problem because it uses sentences including the target word as unlabeled data.

5.4. Future works

In the experiment, the effectiveness of our methods was shown for only the verbs in Japanese Dictionary task of Senseval2. However, the effectiveness is a little. For the nouns in that task, the original k-NN achieves the accuracy 76.83%, but our method the 76.07% in actual. Even using ensemble learning, the accuracy cannot be boosted. The best score of the task is 78.5% for nouns, and 79.8% for verbs. Our scores are quite lower than the best scores.

To improve our method, we believe that the k-NN must be improved. Our proposed semi-unsupervied method needs the k-NN as the base learning method. However, the k-NN is not so good for this task, because the instance is maped the high dimensional vector. In general, k-NN using the high dimensional vectors are apt to suffer the ‘*curse of dimensionality*.’ A way to overcome this problem is to map a high dimensional vector to a low dimensional vector. For example, the PCA (principal component analysis) and the KL expansion (Karhunen-Loève expansion) are availave to do it.

Moreover, we fixed k to 5, but the proper k depends on the problem. We have to estimate proper parameters.

6. Conclusions

In this paper, we proposed the semi-supervised learning method using Fuzzy clustering to solve WSD problems. Moreover, we reduced side effects of semi-supervised learning by ensemble learning. Our method regards a labeled instance as the prototype of a class, and moves it to suitable points by Fuzzy clustering. We can classify a test instance by the k-NN with the moved prototypes. Moreover, to reduce side effects of semi-supervised learning, we used the ensemble learning combined the k-NN with initial labeled data and the k-NN with new labeled data obtained by Fuzzy clustering. In experiments, we took the verb words in Japanese dictionary task of SENSEVAL2. The result showed the proposed method is effective. In future, we have to improve the original k-NN method. One way to do it is to map a high dimensional vector to a low dimensional vector.

7. References

Blum, Avrim and Tom Mitchell, 1998. Combining Labeled and Unlabeled Data with Co-Training. In *11th Annual Conference on Computational Learning Theory (COLT-98)*.

Kurohashi, Sadao and Kiyoaki Shirai, 2001. SENSEVAL-2 Japanese Tasks (in Japansese). In *Technical Report of IEICE*, NLC-36-48.

Miyamoto, Sadaaki, 1999. *Cluster-bunseki-nyuumon (in Japansese)*. MORIKITA Publisher.

Nigam, Kamal, Andrew McCallum, Sebastian Thrun, and Tom Mitchell, 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3):103–134.

Shinnou, Hiroyuki and Minoru Sasaki, 2003. Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the EM algorithm. In *CoNLL-2003: Seventh Conference on Natural Language Learning*.

Ueda, Naonori and Ryohei Nakano, 1997. Analysis of generalization error on ensemble learning (in japanese). *The Institute of Electronics, Information and Communication Engineers*, J80-D2(9):2512–2521.

Yarowsky, David, 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33th Annual Meeting of the Association for Computational Linguistics (ACL-95)*.