Estimating Reliability of Contextual Evidences in Decision-List Classifiers under Bayesian Learning

Yoshimasa Tsuruoka and Takashi Chikayama
School of Engineering and School of Frontier Sciences,
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 JAPAN
tsuruoka, chikayama@logos.t.u-tokyo.ac.jp

Abstract

Classifiers are often required to output not only a classification result but also the probability of the classification. We focus on the decision list classifier which has successfully been applied to a wide variety of NLP tasks. We propose methods based on Bayesian learning to calculate the reliability of contextual evidences in decision lists, which enables decision lists to output theoretically well-founded probabilities. Experimental results obtained using Senseval-1 data set show that our methods enable decision lists to output probabilities appropriately reflecting their reliabilities and improve the classification performance of the decision list algorithm.

1 Introduction

Classifiers are often required to output not only a classification result but also the probability of the classification. Probabilities are required when the overall decision is made by combining outputs of multiple classifications. An example is using a Viterbi search to combine classification results of tokens to find the highest probability path for recognizing named entities (Borthwick et al., 1998). In such cases, accuracy of output probabilities is as important as classification accuracy itself.

Among many supervised classifiers, we focus on the decision list classifier which has been applied to a wide variety of NLP tasks and has shown its effectiveness (Yarowsky, 1994; Hiroyuki, 2000; Usuro and others, 1999). In particular, a decision lists based system is one of the best systems at the word sense disambiguation competition Senseval-1 (Kilgarriff and Rosenzweig, 2000).

However, the majority of research efforts using decision lists do not think much of the output probabilities. We propose a method based on Bayesian learning (Gelman et al., 1995) to calculate the reliability of contextual evidences in decision lists, which enables decision lists to output theoretically well-founded probabilities.

In this paper, we explain and evaluate the proposed method through application to word sense disambiguation problems. Section 2 describes the decision list classifier for word sense disambiguation problems. Section 3 and 4 present the method for estimating the probabilities of contextual evidences based on Bayesian learning and its improvement. Section 5 describes experimental results using Senseval-1 data set. Finally, advantages, limitations and future research directions are discussed in Section 6.

2 The decision-list classifier

The decision list algorithm ranks classification rules by their reliabilities. Classification is made by using the most reliable rule that can be applied to the given context.

Table 1 shows an example of a decision list to disambiguate a polysemous word *plant* (A: living, B: factory). The first rule means that if the word right adjacent to the target is 'life' , the sense is A. The fourth rule means that if the word 'manufacturing' appears within 2-10

Table 1: An example of a decision list (Yarowsky, 1995)

reliability	evidence	sense
8.10	*plant* **life**	A
7.58	**manufacturing** *plant*	B
7.39	**life** (±2–10words)	A
7.20	**manufacturing** (±2–10words)	B
6.27	**animal** (±2–10words)	A
4.70	**equipment** (±2–10words)	B
4.39	**employee** (±2–10words)	B
:	:	:

words from the target, the sense is B.

In this paper, we use the following types of contextual evidences.

- Window

  Word found in ±10 word window

- Word

  Word immediately to the right

  Target word itself

  Word immediately to the left

- Pair of words

  Pair of words at offsets -2 and -1

  Pair of words at offsets -1 and +1

  Pair of words at offsets +1 and +2

The rules are learned from a training corpus. The majority of research efforts adopt the following equation to calculate the reliabililty of each rule.

$$(reliability) = \log\Big(\frac{P(S_j|E_i)}{P(\neg S_j|E_i)}\Big), \quad (1)$$

where S_j is a candidate class for the classification task and E_i is the contextual evidence.

However, to make the decision list algorithm output probabilities, we adopt the following equation as the reliability of a rule.

$$(reliability) = P(S_j|E_i). \quad (2)$$

It should be noted that Equation 1 is a monotonous increasing function of $P(S_j|E_i)$. Since decision lists consider only order of rules, Equation 1 and 2 make equivalent decision lists when probabilities are ideally estimated.

If we have a large number of samples concerning the evidence, $P(S_j|E_i)$ can be easily estimated by *maximum likelihood estimation* for Bernoulli trials:

$$P(S_j|E_i) = \frac{f(S_j, E_i)}{f(E_i)}, \quad (3)$$

where $f(E_i)$ is the number of samples in which E_i occurs and $f(S_j, E_i)$ is the number of samples in which E_i occurs together with S_j.

However, we do not always have sufficient number of samples. For instance, if

$$f(S_j, E_i) = 1, f(E_i) = 1, \quad (4)$$

the probability becomes $1/1 = 100\%$ by Equation 3. This is undesirable estimation, because the decision list gives this kind of infrequent (hence not reliable) evidences the highest priority.

This problem is one of the data sparseness problems and is an essential problem of corpus based methods. We tackle this problem with Bayesian leaning in the next section.

3 Probability estimation based on Bayesian learning

We propose a method based on Bayesian learning for estimating the probability of a rule.

At first, we regard the probability $\theta = P(S_j|E_i)$ as a stochastic variable. The objective is to estimate the expectation value of θ.

Under Bayesian learning, the posterior distribution is given by:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} \quad (5)$$

$$= \frac{P(\theta)P(D|\theta)}{\int_0^1 P(\theta)P(D|\theta)d\theta}, \quad (6)$$

where D is the data we have observed. Since the data D can be regarded as Bernoulli trials in this case, the conditional distribution is:

$$P(D|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}. \quad (7)$$

For simplicity, we denote $f(S_j, E_i)$ by k and $f(E_i)$ by n. Then,

$$P(\theta|D) = \frac{P(\theta)\theta^k(1-\theta)^{n-k}}{\int_0^1 P(\theta)\theta^k(1-\theta)^{n-k}d\theta}. \quad (8)$$

To calculate this posterior distribution, we need to define the prior distribution $P(\theta)$. At this place, we use the uniform distribution which indicates that we have no prior knowledge about the probability distribution of the variable. Then,

$$P(\theta) = 1. \quad (9)$$

The posterior distribution is:

$$P(\theta|D)$$
$$= \frac{\theta^k(1-\theta)^{n-k}}{\int_0^1 \theta^k(1-\theta)^{n-k}d\theta} \quad (10)$$
$$= \frac{\theta^{(k+1)-1}(1-\theta)^{(n+2)-(k+1)-1}}{\int_0^1 \theta^{(k+1)-1}(1-\theta)^{(n+2)-(k+1)-1}d\theta} \quad (11)$$

This probability distribution is called the *beta distribution* and the expectation value is given by:

$$E[\theta] = \frac{k+1}{n+2}. \quad (12)$$

Finally, the reliability of a rule is given by:

$$(reliability) = \frac{k+1}{n+2} = \frac{f(S_j, E_i)+1}{f(E_i)+2}. \quad (13)$$

This equation has a similar form to Equation 3. We can compute reliabilities in almost the same way as maximum likelihood estimation except for using $f(S_j, E_i)+1$ and $f(E_i)+2$ instead of $f(S_j, E_i)$ and $f(E_i)$ respectively.

Our experimental results in Section 5 show that this estimation method enables decision lists to output probabilities which indicate reliability of classifications. However, there are still gaps between expected accuracies computed by averaging output probabilities associated with each classification and actual accuracies obtained. Reasons suspected are the followings.

- Global vs. history-conditional

  Classification by a certain rule implies that no earlier rules have not matched the given context. However, Equation 2 ignores this history-condition and computes probabilities merely from the global frequencies. If the contextual evidence of the rule is independent of those of the earlier rules, the history-conditional probability would be the same as the global probability. However, this is not the case in practice. With regard to this problem, Yarowsky computes probabilities via the interpolation of the global and history-conditional probabilities (Yarowsky, 2000).

- Training data vs. test data

  If the statistical property of the training data is different from the property of the test data, the actual precision suffers. This is a fundamental problem of corpus-based natural language processing.

- Prior distribution

  We have assumed that the prior distribution is uniform. Suppose, however, that there are five candidate classes and we have no information about the distribution, the prior distribution should be the distribution which has its peek around 0.2, while the uniform distribution is not such a distribution. The unifrom distribution is not always appropriate for the prior distribution.

In this paper, we are not concerned with the first two issues. We address the third issue in the next section.

4 Prior distribution

Under Bayesian learning we can compute posterior probability distribution more accurately by using appropriate prior distribution. The question is which probability distribution appropriately reflects our prior knowldge.

Here we make an assumption that the rules with a small number of examples have similar probabilistic property to those with a large

number of examples. This assumption allows us to form a prior distribution from the actual probability values of the rules with a large number of examples. The boxes in Figure 1 show examples of actual relative frequencies of the probability values of the rules which have more than 10 examples.

To make use of these observed probability values as a prior distribution, we adopt the *beta* distribution which can flexibly approximate various shapes of probability distributions by its two parameters. Furthermore, the distribution is the 'natural conjugate prior distribution' for Bernoulli trials, which enables us to compute analytically the posterior distribution without difficulty. The two parameters are set in such a way that the expectation value and variance of the *beta* distribution are made equal to those of the observed probability values[1]. The curves in Figure 1 represent the beta distributions determined in this way.

In order to take into account the differences of probabilistic properties among different types of contextual evidences (Window, Word, Pair of words), we separately conduct the above procedure for each evidence type.

Adopting the *beta* distribution, the prior distribution is represented by:

$$P(\theta) = \frac{1}{B(a,b)}\theta^{(a-1)}(1-\theta)^{(b-1)}, \qquad (14)$$

where $B(a,b)$ is the *beta* function:

$$B(a,b) = \int_0^1 \theta^{(a-1)}(1-\theta)^{(b-1)}. \qquad (15)$$

Substituting this prior distribution into Equation 8, we obtain the posterior distribution:

$$P(\theta|D) = \frac{\theta^{(a+k-1)}(1-\theta)^{(b+n-k-1)}}{B(a+k, b+n-k)}. \qquad (16)$$

The expectation value is given by:

$$E[\theta] = \frac{a+k}{a+b+n}. \qquad (17)$$

[1]This parameter estimation method is called 'moment estimation'

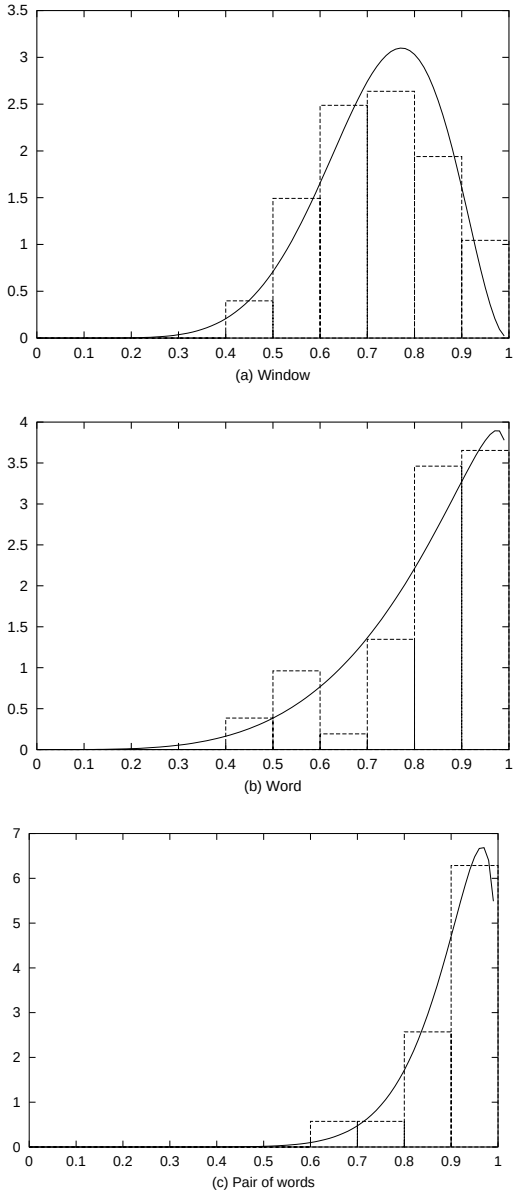

Figure 1: Examples of prior distribution for each contextual type. The target word is 'accident (n).' The boxes represent the histogram of the probability values. Each curve represents the beta distribution which approximates the distribution of those probability values.

Table 2: Actual precisions and expected precisions for Senseval-1 data when the prior distributions is uniform

Word (Pos)	Actual Precision (1)	Expected Precision (2)	$\lvert(1)-(2)\rvert$
accident (n)	83.9%	91.7%	7.8
amaze (v)	100.0%	99.0%	1.0
band (p)	84.1%	90.0%	5.9
behaviour (n)	94.6%	97.4%	2.7
bet (n)	45.1%	72.8%	27.8
bet (v)	70.7%	80.8%	10.1
bitter (p)	49.2%	71.4%	22.2
bother (v)	78.5%	84.0%	5.5
brilliant (a)	47.6%	72.5%	24.9
bury (v)	49.3%	71.6%	22.4
calculate (v)	86.7%	89.3%	2.6
consume (v)	42.6%	71.8%	29.1
derive (v)	54.4%	75.2%	20.8
excess (n)	81.7%	90.6%	8.9
float (n)	53.3%	75.2%	21.8
float (v)	45.0%	72.7%	27.7
floating (a)	59.6%	72.5%	13.0
generous (a)	49.3%	72.9%	23.5
giant (a)	96.9%	95.0%	1.9
giant (n)	79.7%	85.5%	5.9
invade (v)	46.9%	71.3%	24.4
knee (n)	70.9%	79.2%	8.3
modest (a)	66.3%	77.1%	10.8
onion (n)	84.6%	91.1%	6.5
promise (n)	74.3%	81.7%	7.4
promise (v)	88.4%	93.8%	5.4
sack (n)	81.7%	85.0%	3.3
sack (v)	97.8%	98.4%	0.6
sanction (p)	74.5%	82.1%	7.6
scrap (n)	41.7%	80.3%	38.7
scrap (v)	87.6%	89.4%	1.8
seize (v)	60.2%	77.3%	17.1
shake (p)	61.2%	81.0%	19.7
shirt (n)	83.7%	87.4%	3.7
slight (a)	94.0%	94.3%	0.3
wooden (a)	93.9%	97.0%	3.1
Average	71.1%	83.3%	12.3

Table 3: Actual precisions and expected precisions for Senseval-1 data when the prior distribution is the *beta* distribution.

Word (Pos)	Actual Precision (1)	Expected Precision (2)	$\lvert(1)-(2)\rvert$
accident (n)	89.9%	93.1%	3.2
amaze (v)	100.0%	99.8%	0.2
band (p)	87.4%	90.0%	2.6
behaviour (n)	94.6%	97.7%	3.1
bet (n)	50.5%	60.0%	9.5
bet (v)	76.7%	79.0%	2.3
bitter (p)	51.1%	55.1%	4.0
bother (v)	78.0%	83.1%	5.1
brilliant (a)	49.3%	61.0%	11.7
bury (v)	49.3%	59.2%	10.0
calculate (v)	86.7%	89.5%	2.8
consume (v)	42.1%	60.4%	18.3
derive (v)	64.1%	60.6%	3.5
excess (n)	81.2%	90.9%	9.7
float (n)	56.0%	61.3%	5.3
float (v)	42.8%	59.3%	16.5
floating (a)	61.7%	69.4%	7.7
generous (a)	46.7%	56.5%	9.8
giant (a)	96.9%	96.0%	0.9
giant (n)	79.7%	83.8%	4.2
invade (v)	50.2%	65.4%	15.1
knee (n)	72.9%	72.7%	0.2
modest (a)	66.3%	68.8%	2.5
onion (n)	84.6%	94.4%	9.8
promise (n)	74.3%	79.8%	5.4
promise (v)	92.9%	94.5%	1.7
sack (n)	85.4%	82.8%	2.6
sack (v)	97.8%	99.3%	1.5
sanction (p)	77.7%	80.0%	2.3
scrap (n)	45.5%	77.3%	31.8
scrap (v)	87.6%	92.1%	4.5
seize (v)	64.5%	68.8%	4.3
shake (p)	61.2%	78.8%	17.6
shirt (n)	82.6%	85.2%	2.6
slight (a)	94.0%	94.8%	0.8
wooden (a)	93.9%	97.8%	3.9
Average	72.7%	78.8%	6.6

Finally, the reliability of a rule is given by:

$$(reliability) = \frac{f(S_j, E_i) + a}{f(E_i) + a + b}. \qquad (18)$$

5 Experiments

We evaluate our proposed method on Senseval-1 data set which is publicly available online [2]. The data set contains 36 'trainable' polysemous words (for which tagged training data was available). No pre-processing such as stemming, part-of-speech tagging, or parsing has not been conducted.

[2]http://www.itri.brighton.ac.uk/events/senseval/

First, we show the results of the experiments where the prior distribution is uniform. Shown in Table 2 are the actual precisions, the expected precisions and the gaps between them. The expected precision is computed by averaging output probabilities associated with each classification. If each output probability ideally indicates 'true' probability of the classification, the expected precision will be almost equal to the actual precision. Thus, the gaps indicate the goodness of the probability estimation. The smaller the gaps are, the better the estimations are.

Table 4: Comparison of classification performance

	Average Precision
Log-likelihood ratio	71.1%
Thinning out	69.4%
Bayesian (uniform)	71.1%
Bayesian (*beta*)	72.7%

Table 3 shows the results of the experiments where the prior distribution is the *beta* distribution as we described in Section 4. Notice that the average of the gaps has reduced almost by half. Furthermore, the overall classification performance (actual accuracies) has improved.

The proposed methods are not attractive if they deteriorate the classification performance of the decision list algorithm. We conduct experiments to evaluate the classification performance of our methods comparing to some conventional methods. The followings are the conventional methods used in the experiments.

- Log-likelihood ratio

  Reliabilities are calculated by Equation 1. Then,

  $$(reliability) = \log\Big(\frac{f(S_j, E_i) + \alpha}{f(\neg S_j, E_i) + \alpha}\Big). \quad (19)$$

  To avoid zero-denominator, we add a small constant α to the numerator and denominator (Yarowsky, 1994), where α is selected from 0.05, 0.1, 0.2, 0.4, 0.8, 1.6 and 3.2 to optimize classification performance. In this experiment, the best α is 0.8.

- Thinning out

  Reliabilities are calculated by Equation 3. To avoid giving the highest priority to the rules with infrequent evidences. The rules which do not have more than $n_{threshold}$ evidences are thinned out. In this experiment, the best $n_{threshold}$ is 3.

Table 4 shows the classification accuracies where the parameters of the conventional methods are optimized for the data set. Notice that the proposed method with the uniform prior distribution achieves comparative performance to Log-likelihood, despite the fact that it does not require any parameter tuning. The proposed method with *beta* prior distribution achieves the best classification performance[3]

6 Conclusion

To make the decision list algorithm output accurate probabilities, we have proposed a probability estimation method based on Bayesian learning that gives well-founded probability estimations.

Experimental results obtained using Senseval-1 data set show that Bayesian learning with the uniform prior distribution enables decision lists to output probabilities reflecting their reliabilities.

We have also presented a method to make use of prior distributions. The experimental results show that this augmentation significantly reduces the gaps between output probabilities and 'true' probabilities. The results also show that the classification performance of the decision list algorithm is also improved.

The future direction of this study will be to investigate the effectiveness of the proposed methods with other applications than word sense disambiguation problems.

References

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. *In Proceedings of the Sixth Workshop on Very Large Corpora New Brunswick, New Jersey. Association for Computational Linguistics.*

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis.* Chapman & Hall.

[3]This accuracy (72.7%) is considerably lower than the accuracy of the best system in Senseval-1(78%) (Kilgarriff and Rosenzweig, 2000). Preprocessings such as stemming or part-of-speech tagging would improve this accuracy.

Shinnou Hiroyuki. 2000. Conversion of japanese morphological analysis into classification problem and its solving. *IPSJ SIG Notes 2000-NL-135*, 2000(1). (in Japanese).

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and Humanities*, 34(1-2):1–13.

Takehito Usuro et al. 1999. Extraction of preference of dependency between japanese subordinate clauses from corpus and its evaluation. *Journal of Natural Language Processing*, 6(7):29–60. (in Japanese).

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

David Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and Humanities*, pages 34(2):179–186.