Local-Global-Learning of Naive Bayesian Classifier

Shengtong Zhong, Helge Langseth

Department of Computer and Information Science
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
{shket, helgel}@idi.ntnu.no

Abstract

Naive Bayes (NB) models are among the simplest probabilistic classifiers. However, they often perform surprisingly well in practice, even though they are based on the strong assumption that all attributes are conditionally independent given the class variable. The vast majority of research on NB models assume that the conditional probability tables in the model are either learned by maximum likelihood or Bayesian methods, even though it is well documented that learning NB models in this way may harm the expressiveness of the models. In this paper, we focus on an alternative technique for learning the conditional probability tables from data. Instead of frequency counting (which leads to maximum likelihood parameters), we propose a learning method that we call "local-global-learning". We learn the (local) conditional probability tables under the guidance of the (global) NB model learnt thus far. The conditional probabilities learned by local-global-learning are therefore geared towards maximizing the classification accuracy of the models instead of maximizing the likelihood of the training data. We show through extensive experiments that local-global-learning can significantly improve the classification accuracy of NB models when compared to traditional maximum likelihood learning.

1. Introduction

Naive Bayes (NB) models have the strong assumption that all attributes are conditionally independent given the class variable (Figure 1). Even though NB models are among the simplest probabilistic classifiers, they often perform surprisingly well in practice. When learning a NB classifier, the classifier is constructed from a given set of labeled training instances. Assume that training data set D has a number of instances, each instance has n discrete-valued attributes $A_1, A_2, \dots, A_n$ and is represented by a vector $\mathbf{a} = (a_1, a_2, \dots, a_n)$ where a_i is the value of A_i . Let C

denote the class variable with m classes. The NB classifier select the most likely classification of instance $\mathbf{a}$ as:

Figure 1. Naive Bayes models

$$C^*(\mathbf{a}) = \arg \max_c P(C = c) \prod_{i=1}^n P(a_i | c) \quad (1)$$

As all attributes are conditionally independent given the class variable, the classifier is determined by $P(C = c)$, the probability distribution of class variable, and $P(a_i | c)$, the local probability distribution for each attribute. The local distribution is specified by a conditional probability table (CPT). The vast majority of research on NB models assume that the CPT in the model are either learned by maximum likelihood or Bayesian methods.

In general, maximum likelihood inference is simply achieved by frequency counting when learning a CPT for the NB classifier, and thus is a typical generative learning method. Jaeger [5] showed that a NB model can recognize any linearly separable concept, and argues that previous results of representation power for the NB models [2] are due to these authors assuming that all learning are based on the maximum likelihood principle. This indicates that other learning paradigms may lead to better classification accuracy than maximum likelihood learning. Further evidence for this idea is given by Ng & Jordan [4], who showed that discriminative classifiers work better than generative classifiers when there is no missing data, and particularly for small data sets. In classification, the object is to maximize classification accuracy, rather than maximum likelihood. Hence, it may be beneficial to consider learning CPT with discriminative method, as this could improve the

classification accuracy of NB classifier.

In this paper we focus on an alternative technique for learning the CPT. Instead of frequency counting, we propose a discriminative learning method that we call "local-global-learning" (LGL). We learn the (local) CPT under the guidance of the (global) NB classifier learnt thus far. The CPT learned by LGL are therefore geared towards maximizing the classification accuracy of the models instead of maximizing the likelihood of the training data. The proposed LGL method can significantly improve the classification accuracy of NB models compared to traditional maximum likelihood learning. The LGL method is presented in section 2. Section 3 illustrate experiment results which show the effectiveness and efficiency of the proposed method. The conclusion is in the following Section 4.

2. Local-Global-Learning

To explain the difference between traditional Naive Bayes and Local-Global-Learning, the algorithm of learning traditional Naive Bayes classifier with frequency counting is first introduced in Algorithm 1.

Algorithm 1 The traditional Naive Bayes algorithm

```

1: Set all entries of CPT to 0
2: for each observation do
3:   for  $i = 1 : n$  do
4:     Increase  $N(c, a_i)$  by 1
5:   end for
6: end for
7:  $\hat{P}(a_i|c) \leftarrow \frac{N(c, a_i)}{\sum_{a_i} N(c, a_i)}$ 

```

When learning CPT for a traditional NB classifier, the directly affected entries $N(c, a_i)$, where class label and attributes are in observation, are increased by one in frequency counting. Although frequency counting targets to the maximum likelihood of the training dataset, the object of classification problem is to maximize classification accuracy rather than the maximum likelihood. The argument from [4, 5] indicates that learning CPT with the discriminative method could improve the general accuracy of classification of the NB classifier. Instead of learning CPT with frequency counting, we propose a discriminative learning method called "local-global-learning" (LGL). LGL is distinguished from maximum likelihood learning in step 4 of Fig. 1, instead of increasing $N(c, a_i)$ by 1, the contribution of each attribute of the training instance for CPT is guided by the NB classifier, $\hat{C}^*(\mathbf{a})$, learnt thus far.

$$\hat{C}^*(\mathbf{a}) = \arg \max_c \hat{P}(C = c) \prod_{i=1}^n \hat{P}(a_i|c) \quad (2)$$

where $\hat{P}(C = c)$ is the probability of class variables from learnt training instance, $\hat{P}(a_i|c)$ is the probability distribution for each attribute learnt thus far.

Intuitively, the better the current NB classifies a training instance, the less the NB should update its CPT. Informally, this is inline with the old saying "if it ain't broke, don't fix it". Continuing this agreement, one should take the current classifier's ability to classify a training example into account in line 4 of Algorithm 1. If a training example is perfectly classified (100% of the probability allocated to the correct class), the corresponding entries of CPT should be updated by a small amount to strengthen the belief in the current classification; if the training example is misclassified, a comparatively larger amount should be added to the corresponding entries of CPT. This update is denoted by $\Delta N(c, \mathbf{a})$ (compare with the line 4 of Algorithm 1).

In detail, LGL iterates through the training instances. The NB classifier learnt thus far first predict the probability of the current training instance $\mathbf{a}$ in its known true class c .

$$\hat{P}(c|\mathbf{a}) = \hat{P}(C = c) \prod_{i=1}^n \hat{P}(a_i|c) \quad (3)$$

$\hat{P}(c|\mathbf{a})$ reflect the confidence of current NB classifier. Next, the CPT need to be updated based on $\hat{P}(c|\mathbf{a})$. In the experiments reported in section 3, we used $\Delta N(c, \mathbf{a}) = \frac{1}{\hat{P}(c|\mathbf{a})}$ and we call this algorithm $LGL_{\frac{1}{\hat{P}}}$. The exact functional form of this update, $\Delta N(c, \mathbf{a})$, can be debated. However, it seems reasonable that $\Delta N(c, \mathbf{a})$ is a positive and strictly decreasing function of $\hat{P}(c|\mathbf{a})$. This guarantees that:

1. the updating of CPT is under the guidance of global NB classifier learnt thus far, which targets to learn a NB classifier with higher classification accuracy;
2. the updates rule meet with the intuition that the better the instance is classified by the current NB classifier, the less CPT need to learn from the instance and vice versa.

The detailed LGL algorithm is shown in Algorithm 2.

Algorithm 2 Local-Global-Learning

```

1: Set all entries of CPT to 0
2: for each observation do
3:   for  $i = 1 : n$  do
4:     Increase  $N(c, a_i)$  by  $\Delta N(c, \mathbf{a}) = \frac{1}{\hat{P}(c|\mathbf{a})}$ 
5:   end for
6: end for

```

Obviously the updating of local CPT is under the guidance of global classifier which has discriminative element compared to traditional frequency counting, and we would

expect the classifier to outperform the NB in most cases [4, 5]. Jiang et al. [10] proposed a similar discriminative method called Discriminative Frequency Estimate (DFE) to improve the performance of NB classifier. In our notation, DFE can be obtained by choosing $\Delta N(c, \mathbf{a}) = 1 - \hat{P}(c|\mathbf{a})$, and can therefore also be denoted $LGL_{1-\hat{P}}$. At the surface, DFE and our learning algorithm appear fairly similar. However, DFE learning has the drawback that it requires to see the data many times - it needs to iterate repeatedly before eventually converging. This effect was noted in [10], but no exact number of required iterations were given, and also a discussion about why the iterations are needed is missing from [10].

Algorithms of the type we are discussing here have the potential drawback that the ordering of the training examples can influence the classification accuracy. The DFE is particularly vulnerable to this case. In DFE, the corresponding entries of CPT are updated according to $1 - \hat{P}(c|\mathbf{a})$, and the corresponding entries of CPT will not change when the current NB classifier give perfect (100%) classification to a training instance. In practice, the belief of perfect classification need be strengthened. On the contrary, the responding entries are increased even the current training instance is perfectly classified in $LGL_{\frac{1}{P}}$, and each training instance need only to be seen once (which is the same case as in traditional Naive Bayes). Although DFE achieves comparable result as $LGL_{\frac{1}{P}}$, their assumption lead to the more time consuming learning. The detailed discussion appears in next section.

3. Experiment Results

Table 1. Datasets used in experiment

Dataset	#Att	#Cls	Size	Dataset	#Att	#Cls	Size
labor	16	2	57	credit-a	15	2	690
zoo	17	7	101	diabetes	8	2	768
lymph	18	4	148	vehicle	18	4	846
iris	4	3	150	anneal	38	6	898
hepatitis	19	2	155	vowel	13	11	990
autos	25	7	205	credit-g	20	2	1000
sonar	60	2	208	segments	19	7	2310
glass	9	7	214	splice	61	3	3190
audiology	69	24	226	kv-vs-vp	36	2	3196
breast-c	9	2	286	sick	29	2	3772
tumor	17	22	339	hypothyroid	29	4	3772
ionosphere	34	2	351	waveform	40	3	5000
vote	16	2	435	mushroom	22	2	8124
soybean	35	19	683	letter	16	26	20000

In this section the merits of the proposed learning algorithm is investigated by learning classifiers for a number of different datasets (Table 1) taken from the Irvine (UCI) Machine Learning Repository [9]. The datasets are selected

based on their previous usage in similar types of analysis [3, 7, 10].

3.1. Accuracy Results

We have compared the results of the LGL ($LGL_{\frac{1}{P}}$) classifiers to the traditional Naive Bayes (NB) model [1] and the Discriminative Frequency Estimate (DFE) model [10]. The experiments are performed under of framework of WEKA [6]. As these learning algorithms require discrete variables, the attributes were discretized using the unsupervised ten-bin discretization. In addition, instances containing missing attribute-values were removed. The classification results¹ of 10 runs of 10-fold stratified cross validation are shown in table 2. The standard deviation of the accuracy estimate are the theoretical values calculated according to Kohavi [11] and are not necessarily the same as the empirical standard deviations observed during cross-validation.

Table 2. Experiment result on accuracy with stratified cross validation

Dataset	NB	DFE(5)	LGL
anneal	94.37±0.77	97.77±0.49	97.15±0.56
audiology	72.64±2.97	78.49±2.73	79.67±2.68
autos	61.96±3.39	68.97±3.23	69.27±3.22
breast-c	72.70±2.63	70.84±2.69	68.79±2.74
credit-a	84.77±1.37	85.55±1.34	84.93±1.36
credit-g	75.79±1.35	75.46±1.36	74.91±1.37
diabetes	75.77±1.55	75.58±1.55	75.43±1.55
Glass	59.93±3.35	60.66±3.34	61.96±3.32
hepatitis	84.31±2.92	84.72±2.89	85.03±2.87
hypothyroid	92.99±0.42	93.15±0.41	92.95±0.42
ionosphere	90.86±1.54	89.75±1.62	91.09±1.52
iris	94.60±1.85	95.60±1.67	94.80±1.81
kr-vs-kp	87.79±0.58	94.49±0.40	92.88±0.45
labor	92.70±3.45	92.07±3.58	89.93±3.99
letter	70.08±0.32	74.70±0.31	73.54±0.31
lymph	85.97±2.85	85.33±2.91	84.81±2.95
mushroom	95.76±0.22	99.96±0.02	98.74±0.12
tumor	49.71±2.72	48.41±2.71	47.82±2.71
segment	89.05±0.65	90.45±0.61	90.17±0.62
sick	96.57±0.30	97.18±0.27	96.55±0.30
sonar	76.11±2.96	74.15±3.04	76.55±2.94
soybean	92.94±0.98	93.03±0.97	93.81±0.92
splice	95.41±0.37	95.77±0.36	95.64±0.36
vehicle	61.22±1.68	65.12±1.64	66.44±1.62
vote	90.02±1.44	95.33±1.01	93.79±1.16
vowel	65.94±1.51	71.30±1.44	73.57±1.40
waveform	80.02±0.57	80.77±0.56	80.82±0.56
zoo	93.98±2.37	95.05±2.16	95.65±2.03

From the observation of Table 2, LGL achieves significant improvement compared to NB in 9 of 28 datasets

¹Note, we were not able to reproduce the results reported in [10] even though the same experimental settings were used. Results here are in correspondence with results reported elsewhere (e.g., [3, 7, 8]) but sometimes differ from what can be found in [10].

Figure 2. Classification accuracy on subset of audiology dataset with stratified cross validation.

(names given in boldface). Since the model structure of LGL is as simple as traditional Naive Bayes which indicates LGL has the same computation complexity as traditional Naive Bayes, and obviously LGL is an ideal alternative for traditional NB. DFE(5) (iterating training instance five times with DFE) and LGL achieves comparable results in general (despite some wins and losses between each other), however, DFE(5) requires iterating training data many times, and is thus much more time consuming to learn.

In addition, we investigate the influence of the number of the iterations on training instances in DFE and LGL, and we compared the classification accuracy on subset of audiology with stratified cross validation. In Figure 2, the x -axis is the percentage of random chosen subset of audiology dataset, and y -axis is the corresponding classification accuracy. When the training instance is only learnt once, the performance of LGL(1) is much better than DFE(1) and NB. DFE(5) achieves results that are comparable to LGL(1) after 5 iterations over the data. LGL(5) is not much better than LGL(1).

4. Conclusions

In this paper, we propose a novel parameter learning method for Naive Bayes classification. The (local) conditional probability tables are learnt under the guidance of the (global) NB model learnt thus far. The conditional probabilities learnt by local-global-learning are therefore geared towards maximizing the classification accuracy of the models instead of maximizing the likelihood of the training data.

The model of LGL is as simple as Naive Bayes, but the experiment results demonstrate that LGL is efficient and has better classification accuracy than NB in general.

In classification problem, if traditional NB classifier is preferred considering the simplicity of model, LGL is an ideal alternative which has simple model as traditional NB but can achieve better classification accuracy.

In LGL, when receiving a training instance, only the direct affected entries (entries included in training instance) of CPT are updated (the same as NB and DFE). In practice, the NB classifier learnt thus far could also guide the indirect affected entries which in return may improve the classification accuracy. Our next step is to investigate whether an update rule of these indirect affected entries can improve the classification accuracy.

References

- [1] Duda, R. and Hart, P, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
- [2] Domingos, P. and Pazzani, M., "On the optimality of the simple Bayesian classifier under zero-one loss", *Machine Learning*, 29, 103-130, 1997.
- [3] Nir Friedman, Dan Geiger, and Moises Goldszmidt, "Bayesian network classifiers," *Machine Learning*, 29:2-3, 131-163, 1997.
- [4] Andrew Y. Ng and Michael I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," *In: Advances in Neural Information Processing Systems 14*, 841-848, 2001.
- [5] Manfred Jaeger, "Probabilistic classifiers and the concepts they recognize," *In Proceedings of the Twentieth International Conference on Machine Learning*, page 266-273, The AAAI Press, 2003.
- [6] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", *2nd Edition*, Morgan Kaufmann, San Francisco, 2005.
- [7] Helge Langseth and Thomas D. Nielsen, "Classification using Hierarchical Naive Bayes models," *Machine Learning*, 63:2, 135-159, 2006.
- [8] Jiang Su and Harry Zhang, "Full Bayesian network classifiers," *In Proceedings of the 23rd international Conference on Machine Learning*, 897-904, 2006.
- [9] Asuncion, A. and Newman, D.J., "UCI Machine Learning Repository," [<http://www.ics.uci.edu/mllearn/MLRepository.html>], Irvine, CA: University of California, School of Information and Computer Science, 2007.
- [10] Jiang Su, Harry Zhang, Charles X. Ling and Stan Matwin, "Discriminative parameter learning for Bayesian networks," *In Proceedings of the 25th international Conference on Machine Learning*, 1016-1023, 2008.
- [11] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection," *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1137-1143, 1995.