Improved Naive Bayes for Extremely Skewed Misclassification Costs

Aleksander Kołcz[1] and Abdur Chowdhury[1]

AOL, Inc., 44900 Prentice Drive, Dulles VA 20166, USA,
arkolcz,cabdur@aol.com

Abstract. Naive Bayes has been an effective and important classifier in the text categorization domain despite violations of its underlying assumptions. Although quite accurate, it tends to provide poor estimates of the posterior class probabilities, which hampers its application in the cost-sensitive context. The apparent high confidence with which certain errors are made is particularly problematic when misclassification costs are highly skewed, since conservative setting of the decision threshold may greatly decrease the classifier utility. We propose an extension of the Naive Bayes algorithm aiming to discount the confidence with which errors are made. The approach is based on measuring the amount of change to feature distribution necessary to reverse the initial classifier decision and can be implemented efficiently without over-complicating the process of Naive Bayes induction. In experiments with three benchmark document collections, the decision-reversal Naive Bayes is demonstrated to substantially improve over the popular multinomial version of the Naive Bayes algorithm, in some cases performing more than 40% better.

1 Introduction

In certain binary classification problems one is interested in very high precision or very high recall with respect to the target class, especially if the cost of false-positive or false negative misclassifications is disproportionally high. Even though probabilistic cost-sensitive classification frameworks have been proposed, the complicating factor of their successful deployment is uncertainty of precise misclassification costs and the fact that estimation of posterior class probabilities is often inaccurate, especially when dealing with problems involving large numbers of attributes, such as text. As a result, the region within which a classifier can actually benefit the target application may be quite narrow.

In this work, we focus on the problem of extending the utility of the Naive Bayes classifier for problems involving extremely asymmetric misclassification costs. Concentrating on text applications we discuss why certain misclassification errors may be committed with an apparent high confidence and propose an effective method of adjusting the output of Naive Bayes at classification time so as to decrease its overconfidence.

2 Classification with extremely asymmetric misclassification costs

Let us assume a two-class problem $\{(x, y) : y \in \{0, 1\}$ and $x \in \mathcal{X}\}$, where $y = 1$ designates that x belongs to class C (target) and $y = 0$ designates that $x \in \overline{C}$. Assuming no costs associated with making the correct decision, the expected misclassification cost of a classifier F over input domain $\mathcal{X}$ is defined as

$$cost\,(F) = c_{01} P(F = 0 \wedge x \in C) + c_{10} P(F = 1 \wedge x \in \overline{C})$$

where c_{01} is the cost of misclassifying the target as non-target and c_{10} is the cost of making the opposite mistake. If accurate estimates of $P\,(C|x)$ are available, the optimum class assignment for input x results from minimizing the expected loss. In problems with highly asymmetric misclassification costs, assigning x to the more expensive class may be preferable even if its posterior probability is quite low. If $c_{10} \gg c_{01}$, the application dictates very low tolerance for *false positives* and an acceptable classifier needs to be close to 100% correct when assigning objects to class C. Conversely, if $c_{01} \gg c_{10}$ then *false-negatives* are highly penalized and an acceptable classifier needs to be characterized by nearly perfect recall in detecting objects belonging to C. Perfect precision in detecting C is equivalent to perfect recall in detecting $\overline{C}$, but while perfect recall is always possible, perfect precision may not be, especially if the target class is also the one with least examples. In this work we will focus on the problem on achieving near-perfect recall with respect to the target class.

3 Sources of overconfidence in Naive Bayes classification

3.1 The multinomial model

Naive Bayes (NB) is one of the most widely used classifiers, especially in the text domain where it tends to perform quite well, despite the fact that many of its model assumptions are often violated. Several variants of the classifier have been proposed in the literature [1] but in applications involving text, the multinomial model has been found to perform particularly well [2].

Naive Bayesian classifiers impose the assumption of class conditional feature independence which, although rarely valid, has proved to be of surprisingly little significance from the standpoint of classification accuracy [3]. Given input x, NB computes the posterior probability of class C using the Bayes formula $P\,(C|x) = P\,(C)\,\frac{P(x|C)}{P(x)}$. Input x is assigned to the class with the highest expected misclassification cost which, assuming feature independence and when only two classes are present, is determined by the log-odds score:

$$score\,(x) = const + \sum_i \log \frac{P\,(x_i|C)}{P\,\left(x_i|\overline{C}\right)} \qquad (1)$$

3.2 Overconfidence in decision making

It has been recognized that Naive Bayes, while being often surprisingly accurate as a classifier (in terms of the 0/1 loss), tends to be poor when it comes to assessing the confidence of its decisions [4][3]. In particular, the class-probability estimates of NB tend to be clustered near the extreme values of 0 and 1. As shown in [5], this is particularly true in the text domain. When classifying documents with many features, their correlations may compound each other, thus leading to exponential growth in the odds. This effect can intensify in areas only sparsely populated by the training data. Since the log odds in (1) depend on the ratio of class-conditional probabilities, they can be quite high even if the values of the probabilities themselves are very low. But probability estimates for features that were seen relatively rarely in the training data are likely to be more "noisy" than the ones obtained for features with substantial presence. This may result in NB outcomes that appear quite confident even if the neighborhood the test input was only weakly represented in the training set.

Figure 1 illustrates the scatter of NB scores for erroneously classified documents vs. the maximum document frequency (DF) for features contained by these documents. The maximum DF of features in x provides a rough measure of how well the region of containing x was represented by the training data.

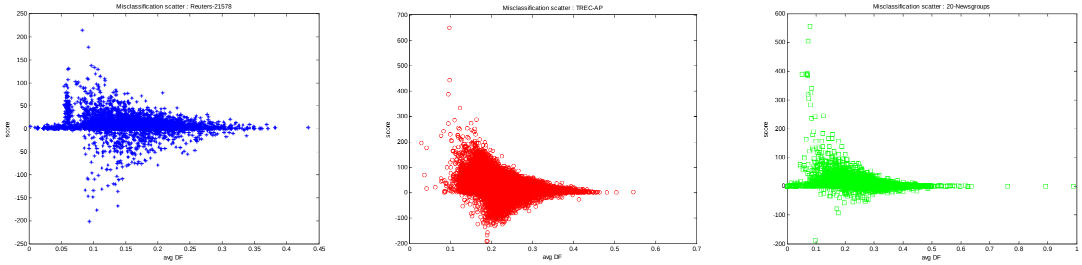

Fig. 1. Scores of Naive Bayes misclassifiications (at default decision threshold) vs. maximum training-set document frequency for features belonging to the misclassified documents (for the collections of: Reuters-21578, 20-Newsgroups and TREC-AP). Misclassifications of documents falling into sparsely populated regions are likely to be made with higher confidence (signified by high absolute score values) than those made for documents for which the training data contained much of related content.

Thus scarcity by itself appears to be a good indicator of overconfidence, although in practice it may be interacting with other factors, such as local class imbalance and document length (e.g., a large number of "noisy" features).

In [6] it was argued that the trust put in the posterior probability estimates of a classifier should decrease with a suitably defined distance between the test input and the training data. In [7] it was suggested that for learners capable of fast incremental learning, the reliability of their posterior estimates can be improved within the framework of transductive learning. Given that a classifier assigns x to class C, it is assumed that a confident decision is one that is little affected by adding x to the training pool of C. With augmented training data, an updated estimate of $P\left(C|x\right)$ is obtained, where its difference to the original is used to gauge the sensitivity of the classifier. The techniques of [6] and [7] both rely on modulating the posterior probability estimate of the base classifier with a normalized reliability indicator, which is interpreted as probability, i.e.,

$$\widehat{P}\left(C|x\right) = P\left(C|x\right) \cdot R\left(C|x\right) \tag{2}$$

where $R\left(C|x\right)$ monotonically approaches 1 as the reliability increases.

4 Changing Naive Bayes' mind: a new reliability measure

The log-odds score of NB has the natural geometric interpretation of the projection of the input onto the weight vector normal to the decision hyperplane. On the other hand, the reliability metrics of [6] and [7], while providing a measure of classifier uncertainty, do not offer a similar interpretation of a margin within which a particular classification is made. We propose a novel reliability metric for Naive Bayes, based on the concept of gauging the difficulty of *reversing* the classification outcome of NB for a given input. Our motivation comes from applying Naive Bayes to on-line learning. Unlike discriminative models such as decision trees, generative learners such as NB can be expected to be stable under small adjustments of the training data. Thus, in order for NB to correct itself, a more extensive change to the distribution of the training data may be needed.

To provide a concrete example, let us consider applying a NB classifier to the problem of spam detection, where a user is given a way to correct classifier mistakes by adding a particular email message to the appropriate pool of training data. Take a scenario where arrival of a spam message finds prompts a corrective action. If an "identical" spam appears again the user responds with another training event, and so on until the classifier correctly identifies the message as spam. In this scenario, the confidence of NB in its initial (mistaken) decision can be linked to the number of training events necessary to correct its outcome, which in turn translates to the amount of change to the training distribution needed for decision reversal. Intuitively, decisions that are confident will require more extensive adjustment of the distribution than less confident ones.

Thus, given that the classifier declares that x belongs to class C, we want to ask how much training with x would it take to reverse its opinion. Since the classifier outcome (1) is determined by its score and assuming the decision threshold of 0 and that the perturbation of the training data does not alter class priors, in order to achieve a decision reversal, one needs to satisfy

$$\log P\left(x|C\right) - \log P\left(x|\overline{C}\right) = \log \widetilde{P}\left(x|\overline{C}\right) - \log \widetilde{P}\left(x|C\right) - \mathit{score} \tag{3}$$

where *score* is the original output score, while $\widetilde{P}\left(x|C\right)$ and $\widetilde{P}\left(x|\overline{C}\right)$ denote estimates over the altered training data.

A question arises as how best to measure the effected change to the training distribution. Here we consider the Kullback-Leibler (KL) divergence, i.e.,

$$rdist(x) = KL\left(P\left(x|\overline{C}\right), \widetilde{P}\left(x|\overline{C}\right)\right) = \sum_{x_i} P\left(x_i|\overline{C}\right) \log \frac{P\left(x_i|\overline{C}\right)}{\widetilde{P}\left(x_i|\overline{C}\right)} \tag{4}$$

Once the KL divergence (4) is computed, a straightforward combination method is to scale (see eq. (2)) the original posterior estimate (for the predicted class) with a suitably defined function of the KL divergence, similarly to the approaches taken in [6] and [7]. Here the difficulty lies in an appropriate choice of the normalization function $R\left(C|x\right) : rdist(x) \rightarrow [0,1]$, but an additional problem with such an approach in the context of Naive Bayes is that the original posterior estimates produced by the NB are already very close to 1 or very close to 0. Thus the modulation of (4) essentially boils down to substituting $R\left(C|x\right)$ for $P\left(C|x\right)$[1]. Given that in the case of extreme misclassification costs one is primarily interested in the narrow region where posterior probabilities are close to 1 or 0, the substitution effect may be undesirable since one loses the original degree-of-confidence information. Therefore, we consider directly modulating the raw log-odds score returned by NB, which typically have a much larger dynamic range:

$$\widehat{score}\left(x\right) = score\left(x\right) \cdot rdist\left(x\right) \tag{5}$$

Other score transformations could be considered. In this work we will also use a function KL distance in the form of:

$$\widehat{score}\left(x\right) = score\left(x\right) \cdot \exp\left(-\gamma \cdot rdist\left(x\right)\right) \tag{6}$$

as an alternative to (5).

5 Experimental Setup

In the experiments described below we compare classifiers at the point where they achieve 100% test-set recall for the target class. At this operating setting, a classifier's utility is measured by its *specificity* (true-negative rate), i.e., the fraction of non-target documents that are classified correctly. Arguably, this measure is very sensitive to class noise and in practice one would have to account for such a possibility, e.g., via interactive or automatic data cleansing procedures.

We compared the proposed *decision-reversal* extension to Naive Bayes (labeled as `NB-KL`) with the following:

- `NB`: Unmodified multinomial Naive Bayes (baseline).
- `NB-Trans`: Kukar's transductive reliability estimator [7] (this is the method closest in spirit to the one proposed here).

[1] In fact [7] does it directly by substituting the posterior estimate of $P\left(C|x\right)$ with $prec \cdot R(C|x)$, where $prec$ refers to the overall precision of the classifier.

```
Algorithm
1. Classify input $x$ using a trained NB model.
2. Estimate the multiplicity $\alpha$ with which $x$ needs to be added to
       the opposite class to achieve decision reversal.
3. Measure the KL divergence (eq.(4)) between the original
       and the perturbed distribution of features for the class
       opposite to the one originally predicted.
4. Modulate the original score (eq.(5) or (6)).
```

Table 1. Steps involved in the decision-reversal Naive Bayes. The most computationally expensive part is step 2, in which one needs to estimate how many corrective events need to take place before the initial decision of the classifier is changed. A naive implementation would keep on generating such events and updating the model, but since in some cases the number of events may be on the order of hundreds or more, this would add significantly to the evaluation time. Instead, we treat the score as a function of the corrective event count a and identify the zero-crossing of score(alpha). In our implementation of the Newton method, usually only 1–7 iterations are needed.

Multi-class problems were treated as a series of two-class tasks, with one class serving as the target and the remaining categories ones as the anti-target, i.e., one-against-the-rest. The results obtained by each classifier and for each dataset are reported by macro-averaging the specificity obtained in the constituent two-class tasks.

5.1 Data sets

We chose three document collections that have often been extensively used in text categorization literature. In each case the collection was split (in the standard way for these collections) into a training set and a test set, which were defined as follows:

- Reuters-21578 (101 categories, 10,724 documents): We used the standard mod_apte split of the data.
- 20 Newsgroups (20 categories, 19,997 documents): A random sample of 2/3 of the dataset was chosen for training with the remaining documents used for testing.
- TREC-AP (20 categories, 209,783 documents): The training/test split described in [8] was used.

Features were extracted by removing markup and punctuation, breaking the documents on whitespace, and converting all characters to lowercase. No stopword removal or stemming was performed. In a modification of the standard bag of words representation, in-document frequencies of terms were ignored.

In all two-class experiments, the feature set was reduced to the top $5,000$ attributes with the highest values of Mutual Information (MI) between the feature variable and the class variable estimated over the training set.

Dataset	NB	NB-Trans	NB-KL	$\frac{\Delta(\text{NB-KL}-\text{NB})}{\text{NB}}$ [%]
Reuters-21578	0.4743	0.3070	**0.6693**	41
20 Newsgroups	0.4033	0.3297	**0.5379**	33
TREC-AP	0.5004	0.1871	**0.5954**	19

Table 2. Macro-averaged classification performance (non-target specificity) captured at the point of perfect target recall. The decision-reversal variant of Naive Bayes consistently outperformed the baseline, while the transductive method consistently underperfomed in all three cases.

6 Results

Table 2 shows the results. For all three datasets, NB-KL provided a substantial improvement over the baseline NB. The transductive method [7] generally underperformed the baseline NB. With hindsight, this is perhaps not too surprising. To achieve high specificity at 100% target class recall, one needs to discount errors for the target class where classification is made with an apparently high confidence. In such cases, the probability of a test document belonging to the target class is estimated by NB to be almost one. The transductive step will increase the probability even further, but this is likely to produce only a very small difference between the original and the final class-probability distributions. Thus the original decision made by NB proves in such cases to be quite stable. It appears therefore that the utility of the transductive method may be highest in cases where the apparent confidence of NB decisions is low.

To examine the effect of an alternative form of score transformation, we evaluated the performance of `NB-KL` using the exponential formula (6) with the choice if γ in $[0.001, 50]$. The best optimization results obtained for `NB-KL` parametrized

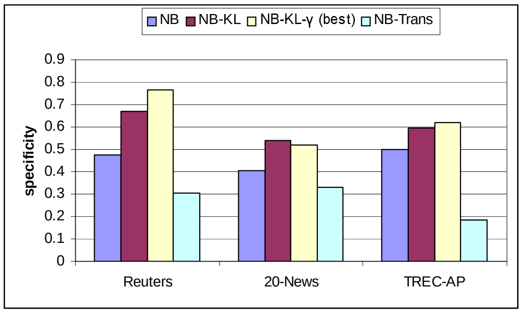

Fig. 2. at the point of perfect target recall. The results for best parameter settings in (6) are compared to the baseline `NB`, `NB-Trans` and the default settings of `NB-KL`. In the case of `Reuters-21578` and `TREC-AP` exponential discounting results in substantial increase in specificity. For `20-Newsgroups`, however, the original formulation of `NB-KL` works better.

according to (6) are compared in Figure 2 with the baseline NB, NB-Trans and the default results for NB-KL. In some cases parameter optimization can substantially improve the performance of NB-KL. The parametric formula (6) was unable however to outperform the regular NB-KL in the case of the 20-Newsgroups dataset. The optimum way of incorporating the decision-reversal information may thus need to be investigated further.

7 Conclusions

The decision-reversal NB proved to be effective in increasing Naive Bayes specificity and countering its native overconfidence. Although the original form of the algorithm performed quite well, further improvements were achieved (in 2 out of 3 datasets) by considering an alternative exponential form of discounting the perturbation distance. The dependence of the effectiveness of incorporating the decision reversal information on the form of the discounting function will be the subject of future work. We are also intending to investigate the effects of combining the proposed method of curbing the overconfidence with techniques motivated by explicit reduction of feature interdependence (e.g., as realized by feature selection).

References

1. Lewis, D.D.: Naive (Bayes) at forty: the independence assumption in information retrieval. In: Proceedings of the 10th European Conference on Machine Learning. (1998) 4–15
2. McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. (1998)
3. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning **29** (1997) 103–130
4. Webb, G., Pazzani, M.: Adjusted probability naive bayesian induction. In: Proceedings of the 11th Australian Joint Conference on Artificial Intelligence. (1998)
5. Bennett, P.N.: Assessing the calibration of Naive Bayes posterior estimates. Technical Report CMU-CS-00-155, Computer Science Department, School of Computer Science, Carnegie Mellon University (2000)
6. Wu, Y.L., Goh, K.S., Li, B., You, H., Chang, E.Y.: The anatomy of a multimodal information filter. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003). (2003) 462–471
7. Kukar, M.: Transductive reliability estimation for medical diagnosis. Artificial Intelligene in Medicine **29** (2003) 81–106
8. Lewis, D.D., Schapire, R.E., Callan, J.P., Papka, R.: Training algorithms for linear text classifiers. In: Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval. (1996) 298–306