BIAS ESTIMATION AND CORRECTION IN A CLASSIFIER USING PRODUCT OF LIKELIHOOD-GAUSSIANS

T. Nagarajan and Douglas O'Shaughnessy

INRS-EMT, University of Quebec
Montreal, Canada
email: raju,dougo@emt.inrs.ca

ABSTRACT

In any classification task the confusion error, in general, is proportional to the number of classes. This is mainly due to sharing of some common attributes (feature vectors) among different classes. This, in many cases, leads to a serious problem, in the sense that, the classifier itself may be biased towards a specific class or a subset of classes. An ideal classifier is not expected to have any such bias. If we assume that, for a given pair of models and their corresponding training data, the log-likelihoods are distributed normally, the bias of any of these models may be visualized in the likelihood-space as an overlap between Gaussian likelihoods of different models (classes). In this paper, we propose a discriminant measure, using a product of Gaussian likelihoods, to estimate the amount of bias. By adjusting the complexity of the models, we show that this bias can be neutralized and a better classification accuracy can be achieved. Presently, the experiments are carried out on the OGI_MLTS telephone speech corpus on a language identification task. The results show that a better classification accuracy can be achieved without any degradation in the performance of any of the individual classes.

Index terms: Pattern classification, Gaussian distribution, Bias.

1. INTRODUCTION

Gaussian mixture modeling (GMM) and hidden Markov modeling (HMM) based techniques have been successfully used in many classification tasks. Using maximum likelihood estimation, the model parameters can be efficiently estimated by maximizing the likelihoods of the training data of a specific class. However, the major weakness of such techniques is that the models are trained in isolation, in the sense that information about other classes in a given task is not considered. Further, the complexities of the models of all the classes are assumed to be same, especially when the amount of training data for all the classes is the same [1]. This may lead to a sub-optimal set of models especially when dealing with extremely confusing classes, which results in increased classification error.

With a given set of classes, and an appropriate feature, the main goal in designing any classifier is to achieve minimum classification error. For a given feature, when two classes are similar in many senses, i.e, if they share many attributes, the classification error between these two may not be reduced. In such cases, at least the misclassification should be symmetrical, i.e., both the classes confuse equally with each other. The worser case is that, in many situations, one of these classes may be biased severely. In a two-class problem, achieving 50% accuracy for both the classes is much better than acheiving 100% accuracy for one class and 0% for the other, even though the overall performance in both the cases are the same. In other words, the classification accuracy for the all the classes should be more or less the same.

In the place of the MLE method for the estimation of model parameters, researchers have tried MMIE-based methods by considering the rest of the models [2] or a sub-set of the most confusing models [3] [4]. Instead of modifying the parameter-estimation method, an interesting method is presented in [5], in which the data points which do not fit the models well, in other words, the outliers, are removed or de-emphasized. The authors of [5] have proposed another technique, in which the decision boundary, in the likelihood space, between a HMM pair is adjusted to reduce the bias in any of the models [6].

In this paper, we propose a method that is similar to the technique presented in [6], in the sense that the bias removal is carried out in the likelihood space. However, instead of modifying the decision boundary, we adjust the log-likelihoods, by modifying the topology of the corresponding model in such a way that the bias between a pair of HMMs is reduced. If we assume that, for a given pair of models and their corresponding training data, the log-likelihoods are distributed normally[1], the bias of any of these models may be visualized in the likelihood-space as an overlap between Gaussian likelihoods of different models (classes). If a feature distribution is assumed to be a univariate Gaussian, the likelihood distribution may be considered as exponential. However, for n-dimensional features, if the feature distribution is a mixture of Gaussians, the resultant likelihood distribution cannot be easily predicted. Based on the empirical observations (especially when the number of training examples is very high) and for analysis, it is presently assumed as Gaussian distribution.

Previously, we have defined a quantitative measure for the amount of overlap between two Gaussians [7]. In [7], this measure is used, in a continuous speech recognition system, to optimize the topology of the syllable models by considering whether a given model can

[1] If a feature distribution is assumed to be a univariate Gaussian, the likelihood distribution may be considered as exponential. However, for n-dimensional features, if the feature distribution is a mixture of Gaussians, the resultant likelihood distribution cannot be easily predicted. Based on the empirical observations (especially when the number of training examples is very high) and for analysis, it is presently assumed as Gaussian distribution.

discriminate training data of confusing classes from its own. The same measure is used in this work to estimate the amount of bias in a model and to remove it. In the present study, we apply the proposed technique on a language identification task and compare its performance with that of the conventional GMM-based classifier [1]. We show that the classification accuracy can be increased considerably by reducing the bias.

The outline of this paper is as follows. In the next section, we define the bias in any model and a quantitative measure to estimate the bias. The step-by-step procedure used for the bias removal is presented in section 3. The experimental setup and the performance on a language identification task is presented in section 4. Finally, other possible methods to reduce the bias and future directions are discussed in section 5.

2. BIAS ESTIMATION

As mentioned earlier, when two different classes share some attributes in common, the confusion error cannot be avoided. In such cases, the confusion between the two classes is either (a) symmetrical, in the sense that both the classes are confused with each other equally or (b) asymmetrical, i.e., one of the classes is biased. In the present work, we concentrate on handling the second case, where one class is dominated by the other. In regression and classification problems, for a specific class, the bias is generally defined in terms of the error rate [8] [9]. When the models are trained with a reasonable number of parameters (components) the error rate, especially on training data, does not provide the required information about the bias.

Let us consider the feature vectors of two different classes (C_i and C_j) as x_k^i and x_k^j. Let λ_i and λ_j be the models of the classes, C_i and C_j, respectively. Let the likelihoods of the feature vectors of the class C_i for the given models λ_i and λ_j be $p(x_k^i|\lambda_i)$ and $p(x_k^i|\lambda_j)$ respectively. We can assume that these likelihoods are distributed normally in likelihood space with suitable parameters. Let these two Gaussians be $N_{ii}(\mu_{ii}, \sigma_{ii}^2)$ and $N_{ji}(\mu_{ji}, \sigma_{ji}^2)$. Similarly, for the feature vectors of the class C_j, the likelihood-Gaussians are $N_{jj}(\mu_{ii}, \sigma_{jj}^2)$ and $N_{ij}(\mu_{ji}, \sigma_{ij}^2)$. Under ideal conditions, the classification error can be related to the separation between the two Gaussian likelihoods shown in figure 1 (a) and (b). If the overlap between the two Gaussians N_{ii} and N_{ji}, is equal to the overlap between N_{jj} and N_{ij}, then there exists no bias.

Earlier, we have proposed a measure to quantify the amount of overlap between two Gaussians [7]. The same measure is used here also, however, in a different context. The details are given below, with required changes, for clarity purposes. Let $N_{ii}(\mu_{ii}, \sigma_{ii}^2)$ and $N_{ij}(\mu_{ij}, \sigma_{ij}^2)$ be

$$N_{ii} = f[(p(x_k^i|\lambda_i)] = \frac{1}{\sqrt{2\pi}\sigma_{ii}} \; e^{-\frac{(p(x_k^i|\lambda_i) - \mu_{ii})^2}{2\sigma_{ii}^2}}, \quad (1)$$

$$N_{ji} = f[(p(x_k^i|\lambda_j)] = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \; e^{-\frac{(p(x_k^i|\lambda_j) - \mu_{ji})^2}{2\sigma_{ji}^2}}. \quad (2)$$

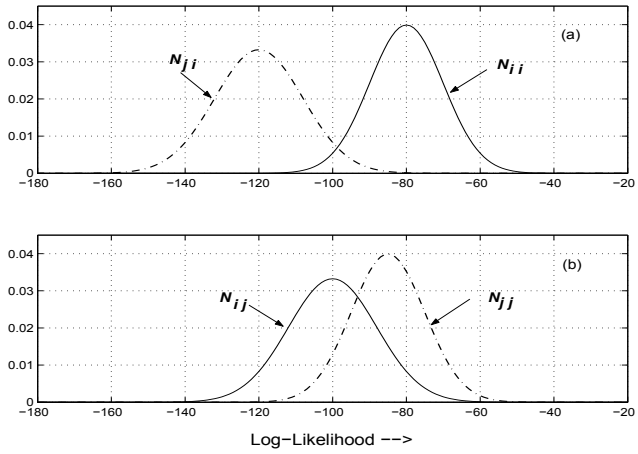

Fig. 1. *An illustration of bias of a model (λ_i) (a) The likelihood distributions (N_{ii} and N_{ji}) of the utterances of the classes C_i for the given models λ_i and λ_j. (b) The likelihood distributions (N_{jj} and N_{ij}) of the utterances of the classes C_j for the given models λ_j and λ_i.*

Let $N_k(\mu_k, \sigma_k^2)$ be[2]

$$N_k(\mu_k, \sigma_k^2) = N_{ii}(\mu_{ii}, \sigma_{ii}^2) \; . \; N_{ij}(\mu_{ji}, \sigma_{ji}^2). \quad (3)$$

For the product of the Gaussians, the mean (μ_k) and its variance (σ_k^2) can be given as

$$\mu_k = \frac{\sigma_{ij}^2 \mu_{ii} + \sigma_{ii}^2 \mu_{ji}}{\sigma_{ii}^2 + \sigma_{ji}^2}, \quad (4)$$

$$\sigma_k^2 = \frac{\sigma_{ii}^2 \sigma_{ji}^2}{\sigma_{ii}^2 + \sigma_{ji}^2}. \quad (5)$$

In order to quantify the amount of overlap between two different Gaussians, we define the following ratio ($\mathcal{O}_{ij}$).

$$\mathcal{O}_{ij} = \frac{\max[N_{ii}(\mu_{ii}, \sigma_{ii}^2) \; . \; N_{ji}(\mu_{ji}, \sigma_{ji}^2)]}{\max[N_{ii}(\mu_{ii}, \sigma_{ii}^2) \; . \; N_{ii}(\mu_{ii}, \sigma_{ii}^2)]} = \frac{Nr}{Dr}. \quad (6)$$

In Equation (6),

$$Nr = \frac{1}{2\pi\sigma_{ii}\sigma_{ji}} \; e^{-\left[\frac{(\mu_k - \mu_{ii})^2}{2\sigma_{ii}^2} + \frac{(\mu_k - \mu_{ji})^2}{2\sigma_{ji}^2}\right]} \quad (7)$$

$$\text{and } \; Dr = \frac{1}{2\pi\sigma_{ii}^2}. \quad (8)$$

From Equations (7) and (8), Equation (6) can be written as

$$\mathcal{O}_{ij} = \frac{\sigma_{ii}}{\sigma_{ji}} \; e^{-\left[\frac{(\mu_k - \mu_{ii})^2}{2\sigma_{ii}^2} + \frac{(\mu_k - \mu_{ji})^2}{2\sigma_{ji}^2}\right]}. \quad (9)$$

If $\mu_{ii} = \mu_{ji}$, then Equation (6) reduces to

$$\mathcal{O}_{ij} = \frac{\sigma_{ii}}{\sigma_{ji}}. \quad (10)$$

[2] In the present study, N_k is not normalized, as this will not affect its use in Equation (11).

However, for this case we expect the overlap $\mathcal{O}_{ij}$ to be equal to 1. To achieve this, Equation (6) is further normalized as given below:

$$\begin{aligned} \mathcal{O}_{ij}^{\mathcal{N}} &= \mathcal{O}_{ij}\frac{\sigma_{ji}}{\sigma_{ii}} \\ &= e^{-\left[\frac{(\mu_k-\mu_{ii})^2}{2\sigma_{ii}^2}+\frac{(\mu_k-\mu_{ji})^2}{2\sigma_{ji}^2}\right]}. \end{aligned} \quad (11)$$

The resultant $\mathcal{O}_{ij}^{\mathcal{N}}$ is used as a measure to estimate the amount of overlap between two Gaussians.

Using this definition for the amount of overlap, the bias (B) can be defined as

$$B = \mathcal{O}_{ij}^{\mathcal{N}} - \mathcal{O}_{ji}^{\mathcal{N}}. \quad (12)$$

If B is positive, the model λ_i is considered as "negatively-biased"; else λ_i is considered as "positively-biased". For each pair of classes, the estimated bias, B, is removed as explained in the next section.

3. BIAS REMOVAL

Let us consider N different classes, $C_1, C_2, ..., C_N$. Let λ_i^m be the acoustic models of the class C_i, where m is the number of mixtures per state, that varies from $1, 2, ..., M$. For each class, M models with varying numbers of mixtures are pre-generated. For the N classes, the number of possible pairs is $N(N-1)/2$. For each pair, the bias is estimated and corrected as given below.

1. For each pair (C_i, C_j), compute the overlaps $\mathcal{O}_{ij}^{\mathcal{N}}$ and $\mathcal{O}_{ji}^{\mathcal{N}}$ (models with M mixtures), using their corresponding training data.
2. Decide the category of the bias (positive or negative) in each of the models.
3. Reduce the number of mixtures of the positively-biased model by 1 ($m = m - 1$). In fact, this can be accomplished by increasing number of mixtures of the negatively-biased model[3]. Increasing the number of mixtures, in other words splitting the mixtures, leads to lower variances and increased likelihoods for the training data. Decreasing the number of mixtures does the reverse. In fact, if we change the number of mixtures of one model, the influence can be seen in all the four Gaussians shown in figure 1.
4. Calculate the new bias as given in equation 12.
5. If there is a sign change in bias or if it is zero, terminate the process, and consider the corresponding number of mixtures for the positively-biased model.
6. Repeat the steps 3-5, otherwise.

In this procedure, in a pair, only one model is modified (specifically a positively-biased model) to remove the bias. However, it can be corrected by modifying (increasing/decreasing the number of mixtures of the negatively/positively-biased models) both the models' complexity simultaneously.

In the current study, this bias-removal technique is used for a language identification task as described in the next section. However, with minor changes, the same technique can be used for any classification task that is based on GMM or HMM.

[3] As mentioned here, the bias can be removed also by increasing the complexity of the negatively-biased model. However, there exists a problem that the resultant model may be over-trained.

4. LANGUAGE IDENTIFICATION TASK

In the spoken language identification task, it should be assumed that no test speaker's spectral (or any other type of) information is present in the training set. In that, the comparison between the test utterance and the reference models of the languages is from unconstrained utterances of two different speakers.

The Oregon Graduate Institute Multi-language Telephone Speech (OGI_MLTS) Corpus [10], which is designed specifically for LID research, is used for both training and testing. This corpus currently consists of spontaneous utterances in 11 languages: English (En), Farsi (Fa), French (Fr), German (Ge), Hindi (Hi), Japanese (Ja), Korean (Ko), Mandarin (Ma), Spanish (Sp), Tamil (Ta) and Vietnamese(Vi). The utterances were produced by $\sim$90 males and $\sim$40 females, in each language over real telephone lines. In our work, presently all 11 languages are used. To maintain the homogeneity in training and testing across languages, for each language the first 30 male speakers' 45 s utterances are used for training and the rest of the male speakers' 45 s utterances are used for testing. The total number of test utterances is 581.

The models are trained and tested using HTK. Cepstral mean subtracted MFCC (13 static + 13 dynamic + 13 acceleration) is used as a feature for this task. Instead of GMM, single-state HMMs with varying number of mixtures (16 to 64, in steps of 1) are trained for all the languages. For each of the pairs, the bias is estimated and removed as explained in the previous section. Here, removing the bias from one model, in a pair, may affect the other classes undesirably. To avoid this, for this language identification task, we adopt a two-level identification as given below.

The proposed technique for bias estimation and correction, in its present form, can be used for pair-wise testing only. However, we have extended this technique for 11 language testing by performing identification in two levels. In the first-level identification, conventional testing is performed using a fixed number of mixtures (here, 64) for all the classes (refer to the third column of Table 1). The 2-best output of this level is considered as the pair for the second-level, in which the bias removed models are used (refer to fourth column of Table 1).

We have compared the performance of the proposed technique with that of a conventional GMM-based classification system, where same complexity is used for all the models (classes) invariably. From the comparison on the performances (refer to Table 1), the following observations can be made.

- The overall performance of the language identification system is improved by 4.2% over the baseline system's performance.
- The performances of the weaker classes are considerably increased (e.g., Japanese).
- Interestingly, the bias-removal technique does not reduce the performance of any of the languages. (Only a minor reduction is observed for English).
- We observe a 2% reduction in standard deviation (refer to last row of the Table 1) also.

Recent developments in language identification tasks show a very low classification error-rate. Our intention, here, is to show that the performance of a GMM/HMM-based system can be improved by

Table 1. Language identification performance of the systems before and after bias removal (BR)

Language	No. of tests	Performance in %	
		Before BR	After BR
En	99	71.7	70.7
Fa	50	56.0	56.0
Fr	49	67.3	71.4
Ge	29	48.3	55.2
Hi	131	51.9	54.9
Ja	23	30.4	47.8
Ko	33	51.5	54.5
Ma	32	62.5	62.5
Sp	43	51.2	53.5
Ta	59	50.8	52.5
Vi	33	42.4	51.5
	Average	53.1	57.3
	std	11.4	9.4

the proposed logic. Since the numbers of female speakers for many of the languages are very small, we have considered only the male speakers' data.

5. DISCUSSION

As mentioned in the introductory part of this paper, the defined-bias can be removed in multiple ways. In the present study, we have adjusted only the complexity of the biased model to remove the bias. The common attributes between classes, either the common feature vectors or at least the common examples (here, a speaker of one class may fully resemble another class), can be removed or de-emphasized to remove bias.

In the present study, by performing identification tasks in two-levels, the technique is extended to a multi-class problem. However, when the numbers of classes are very high, the training process may be expensive. In such cases, for each class, instead of considering all the other classes as competing classes, we can consider only the most confusing class(es) alone. If we assume that for each class there is only one competing class, which in many cases is correct, we can perform classification in a single-level itself.

An interesting point to note here is, the bias is made close-to-zero by keeping the complexity of one system fixed and modifying the other, which makes the problem simpler. If we decide to modify both the models, we may get different points where the bias is close-to-zero. In such a situation, we have to make sure that the absolute overlaps, $\mathcal{O}_{ij}^{\mathcal{N}}$ and $\mathcal{O}_{ji}^{\mathcal{N}}$, between the Gaussians are also reduced. Bias removed models will ensure reduced variance in the performance. However, reduced overlaps between the Gaussians will ensure a better overall performance. An ideal classifier can be realized if both the conditions are satisfied.

6. CONCLUSION

In this paper, we proposed a new technique to estimate and remove bias, if any, in a classifier to improve the classification accuracy of a system. Assuming that the acoustic likelihoods are distributed normally, the bias is estimated in the likelihood-space. The bias term is defined in terms of a product of likelihood-Gaussians. We suggested various methods to remove this bias, and shown that it can be accomplished by modifying the complexity of a model itself. Further, through our experiments on a language identification task, we have shown that the performance of weaker classes can be improved without any significant degradation in the performance of the other classes. Since the proposed bias-removal method is based on likelihoods, it can be utilized in any of the GMM/HMM-based classifiers.

7. REFERENCES

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech, and Audio Processing*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[2] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Tokyo, Japan, 1986, vol. 1, pp. 49–52.

[3] K. Markov, S. Nakagawa, and S. Nakamura, "Discriminative training of hmm using maximum normalized likelihood algorithm," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May 2001, vol. 1, pp. 497–500.

[4] J. K. Chen and F. K. Soong, "An n-best candidates-based discriminative training for speech recognition applications," *IEEE Trans. Speech, and Audio Processing*, vol. 2, no. 1, pp. 206–216, Jan. 1994.

[5] L. M. Arslan and J. H. L. Hansen, "Selective training for hidden Markov models with applications to speech classification," *IEEE Trans. Speech, and Audio Processing*, vol. 7, no. 1, pp. 46–54, Jan. 1999.

[6] L. M. Arslan and J. H. L. Hansen, "Likelihood decision boundary estimation between HMM pairs in speech recognition," *IEEE Trans. Speech, and Audio Processing*, vol. 6, no. 4, pp. 410–414, July 1998.

[7] T. Nagarajan and Douglas O'Shaughnessy, "Discriminative MLE training using a product of Gaussian likelihoods," in *INTERSPEECH - 2006*, Pittsburgh, Pensylvania, USA, Sept. 2006, pp. 601–604.

[8] E. B. Kong and T. G. Dietterich, "Error-correcting output coding corrects bias and variance," in *12th International conference on Machine learning*, 1995, pp. 313–321.

[9] G. M. James, "Variance and bias for general loss function," *Machine Learning*, , no. 51, pp. 115–135, 2003.

[10] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multilanguage telephone speech corpus," in *Proceedings of Int. Conf. Spoken Language Processing*, Oct 1992, pp. 895–898.