TOWARDS LINK CHARACTERIZATION FROM CONTENT

John Grothendieck
Rutgers University

Allen Gorin
U.S. Department of Defense

ABSTRACT

In processing large volumes of speech and language data, we are often interested in the distribution of languages, speakers, topics, etc. For large data sets, these distributions are typically estimated at a given point in time using pattern classification technology. Such estimates can be highly biased, especially for rare classes. While these biases have been addressed in some applications, they have thus far been ignored in the speech and language literature. This neglect causes significant error for low-frequency classes. Correcting this biased distribution involves exploiting uncertain knowledge of the classifier error patterns. The Metropolis-Hastings algorithm allows us to construct a Bayes estimator for the true class proportions. We experimentally evaluate this algorithm for a speaker recognition task. In this experiment, the Bayes estimator reduces maximum RMSE by a factor of five. Performance is furthermore more consistent, with range of RMSE reduced by a factor of 4.

Index Terms— knowledge acquisition, Monte Carlo methods, speech processing

1. INTRODUCTION

There is increasing interest in characterizing links in a communication network, not simply in terms of message count but by content. For example, what proportion of internet traffic is peer-to-peer? There may be little or no prior knowledge. For communication between humans, characterization can involve any of the standard tasks in language processing. We might assign a categorical label (e.g. language, speaker or topic) to linguistic content encoded in audio, text or document images, then focus on the distribution over these categories. Histograms of these distributions provide useful summary statistics to help humans cope with information overload [1].

Automated classifiers have many uses, but their output is typically biased due to classification errors. Proportional bias increases as the frequency of a class decreases. For example, consider some binary task with 5% false alarm rate and negligible missed detections. If 20% of the data is truly from the target class, around 24% of the data will be hypothesized as such by the classifier due to false alarms. This is incorrect, but perhaps still useful. However, for a true value of 0.01%, the expected 5% hypothesized proportion is wrong by orders of magnitude. This large proportional bias is unsatisfactory, especially in applications where rare events are of interest.

Given the classifier error rates, it is straightforward to estimate the most likely class proportions via the E-M algorithm. These can be estimated from some sample set with manual annotation. However, estimates based upon finite data have some degree of uncertainty. Optimal decisions can require understanding of variance — the most likely target class proportion may be 20%, but how plausible is 19%, or 10%? This is a well understood problem in statistics, given the assumption that the test data is drawn from the same population as the training data [2]. To provide variance information rather than a simple point estimate requires a different technical approach.

A hierarchical Bayes model for the true class proportions can incorporate error rate uncertainty. The Metropolis-Hastings (M-H) algorithm [3] allows us to construct the posterior distribution of true class proportions. The posterior mean provides a Bayes estimate of the class proportions, while posterior variance provides confidence bounds on the estimated proportion.

2. RELATED RESEARCH

Issues of data summarization when using a classifier have not been a traditional focus of Human Language Technology (HLT) research. An appropriate model for classifier errors is presented in [2]; this work however does not address the issue of estimation based upon uncertain error rates. The bias inherent in hard classifier output has been ignored by the speech and language processing community (thus such works such as [4] analyze output label rather than true class proportions). The work [5] also seeks a methodology that is valid for all possible class proportions; it further provides an HLT engineer's sketch of Bayesian decision theory. Their interest however is on calibration of score likelihoods conditional on class, analogous to our confusion matrix, rather than updating the hypothesis prior distribution. Natural language processing uses complex classifiers and machine learning techniques, but corpus summary statistics have not been a primary concern.

Research areas involving high-speed high-volume data streams (such as internet traffic) focus more on issues of speed and scalability. Recently there has been convergence with HLT. Content mining techniques are increasingly used to monitor networks [6], while there is ongoing research on fast language processing scalable to massive data streams. [7] describes one application in (text-based) language and topic identification. As high-volume data processing incorporates imperfect classifiers, classifier bias can seriously impact data analysis.

The medical literature recognizes the issue of classification bias; some authors use confusion matrix inversion, assuming known error rates [8]. A few works note that this is unrealistic [9]. In particular, the technical approach of [10] is very similar to ours. Their paper considers only two classes and relies on a Gibbs sampling scheme dependent on conjugate priors, but is readily extensible to more general classification problems. These results seem to be unknown outside of the epidemiology literature.

Our technical problem requires deducing true class proportions from the classifier's hypothesized proportions and estimated error patterns. From this perspective our solution simply adapts standard Bayesian techniques to a particular mixture problem. The justification for using Markov Chain Monte Carlo (MCMC) numerical estimation is well-understood [3], but the practice involves some art [11] [12].

Email: grothend@stat.rutgers.edu
Email: a.gorin@ieee.org

3. ESTIMATING CLASS PROPORTIONS

3.1. Introduction

We measure estimator performance via mean squared error (MSE). In this section, we show that hypothesized class proportions act as a shrinkage estimator towards the fixed eigenvector of the classifier error rate matrix. This introduces uncontrolled bias and lack of predictability into the MSE.

In incorporating a model of error rates, the Bayes estimator described in this paper gains some desirable statistical properties. It is *consistent* in the sense that given unlimited data it must converge to the truth. It is *admissible* (no strictly lower risk estimator exists) since it is Bayesian for a particular prior [13]. By construction it has minimum expected squared error loss under explicit prior beliefs about the parameters.

3.2. The Distribution of Classifier Hypotheses

Denote by x the value of the true class label for some observation, and by y the hypothesized class label from the classifier output. Assume multinomial samples x and y with associated class probability vectors V and W respectively, where $V \equiv \{v_i = P(x = i)\}$ and $W \equiv \{w_i = P(y = i)\}$. Improved classifier performance brings W closer to true V, but accurate estimation of V is possible for imperfect classifiers given accurate knowledge of classifier error rates.

For a given data set and classifier we have a model with class-conditional error probabilities

$$c_{ij} = P(y = i \,|\, x = j)$$

The c_{ij} are independent of the (unknown) true distribution of x. Given probability vectors V for the true-class distribution and W for the hypothesized-class distribution, this leads to the multinomial parameter equation

$$W = CV \tag{1}$$

The matrix C has an eigenvalue 1, thus at least one 'fixed' eigenvector V_F such that $CV_F = V_F$. This V_F is unique so long as the Markov process defined by transitions C is *ergodic* (irreducible with recurrent aperiodic states). A sufficient (though not necessary) condition is if no entries of C are zero. In such a case, V_F is the unique attractor for all probability vectors V under the action of C: $\lim_{n\to\infty} C^n V = V_F$. This creates the bias in hypothesized versus true class proportions — other vectors V are drawn towards V_F. Thus, $W = CV$ differs from V except at V_F.

Given a set Y of observed classifier output, we denote by $y(i)$ the classifier hypothesis for observation i, where $y(i) \in \{1, K\}$ for a classifier with K categories. Denote the number of observations in Y by N_Y. We abuse notation and let Y further denote the vector of hypothesized class counts, so $Y \sim \text{Multi}(N_Y, W)$. Given Y, we have $\hat{W} = Y/N_Y$ the relative frequency estimator for W. Thus $\hat{W}$ is a random variable, while $\hat{W}(Y)$ is a fixed value. The expected MSE of $\hat{W}$ as an estimator of true proportions V has the classic decomposition:

$$E\left[(\hat{W} - V)^2\right] = \text{var}(\hat{W}) + \left[E(\hat{W} - V)\right]^2 \tag{2}$$

Consider the 2-class case. When the number of observations N_Y is large, then $\text{var}(\hat{w}_1)$ is small, squared bias dominates the MSE, and the root mean squared error (RMSE) of $w_1 \approx |E(w_1) - v_1|$. For smaller N_Y, $\text{var}(\hat{w}_1)$ contributes to RMSE. Figure 1 shows an example. Estimator $\hat{w}_1$ suffers from uncontrolled bias due to the shrinkage of V towards V_F. The shrinkage depends on both C and unknown V, so RMSE cannot be predicted without an explicit model for C. Compensating for the bias by estimating C provides a more predictable RMSE.

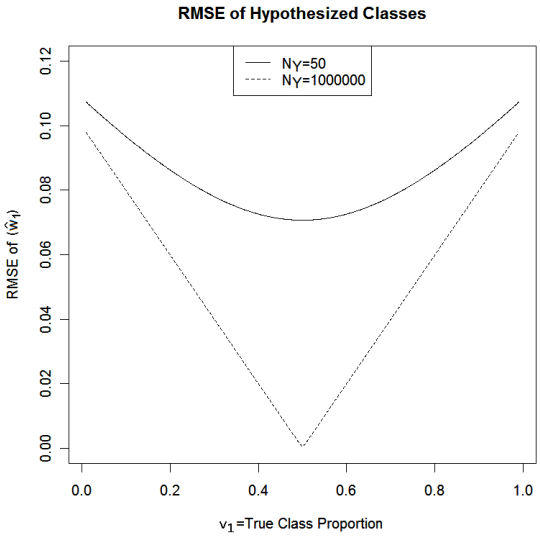

Fig. 1. RMSE of $\hat{w}_1$ for 10% EER classifier, high and low variance.

3.3. Hierarchical Bayes Model

Estimation of error rates C is typically done from some manually labeled corpus L, with l_{ij} the number of observations with true class j and hypothesized class i. The distribution of the parameter V depends on the distributions of W and C, which in turn depend on L and Y. A hierarchical Bayes model can exploit priors not only on the parameter of interest, but on the other parameters on which its distribution depends.

We model true class and class-conditional output labels as multinomial random variables. Flat priors allow us to model $P(W|Y)$ as a Dirichlet and $P(C|L)$ as a hyper-Dirichlet distribution. Joint distribution $P(C, W|L, Y)$ is more complicated in that the domain of W depends on C. Changing coordinates to $P(C, V|L, Y)$ eliminates that issue, but data Y provides information on CV rather than directly on V. Thus we construct posterior $P(C, V|L, Y)$ via random sampling.

3.4. Metropolis-Hastings Estimation of V

Our goal is to estimate the distribution $P(V|L, Y)$, where V is the vector of true class proportions given data L and Y. We have no analytic solution for $P(V|L, Y)$, but do have:

$$P(C, V|L, Y) \propto P_0(C, V)P(Y|W = CV)P(L|C)$$

for prior $P_0(C, V)$ by Bayes Rule. We generate random samples of C and V according to probabilities $P(C|V, L, Y)$ and $P(V|C, Y)$. We recover $P(V|L, Y)$ by projecting onto the marginal distribution.

The M-H algorithm provides a Monte Carlo method for generating samples that are provable convergent to a target distribution. Denote some parameter space by X and the (computable) probability distribution by $q(x)$. M-H performs a random walk in X via a

transition kernel $\pi(x, x')$. The transition kernel defines a Markov chain, which under suitable conditions (i.e. ergodicity) is guaranteed to converge in probability to the target distribution $q(x)$. See [3] and [14] for more details.

We generate a correlated sample of size T as follows:

1. Set initial $C_0 = \hat{C}(L)$, $V_0 = C_0^{-1}\hat{W}(Y)$.
2. For t in 1 to T:
   (a) Select candidate C' via independent transition kernel $\pi_C = P(C'|L)$.
   Define $W'_C = C'V_{t-1}$ and $W_{c,t-1} = C_{t-1}V_{t-1}$
   Thus $\alpha_C = P(W'_C|Y, L)/P(W_{c,t-1}|Y, L)$
   Accept $C_t = C'$ with probability $\min(\alpha_C, 1)$.
   (b) Select candidate V' via the transition π_V.
   Define $W'_V = C_tV'$ and $W_{v,t-1} = C_tV_{t-1}$
   Thus $\alpha_V = P(W'_V|Y, L)/P(W_{v,t-1}|Y, L)$
   Accept $V_t = V'$ with probability $\min(\alpha_V, 1)$.

This random walk in (C, V) will converge to $P(C, V|Y, L)$. Sequence V_t is guaranteed to converge to the marginal distribution of interest, $P(V|Y, L)$. Given K classes this is $O(K^2T)$; the number of classes that can be considered in practice is limited by the amount of labeled data L to estimate C rather than algorithmic complexity.

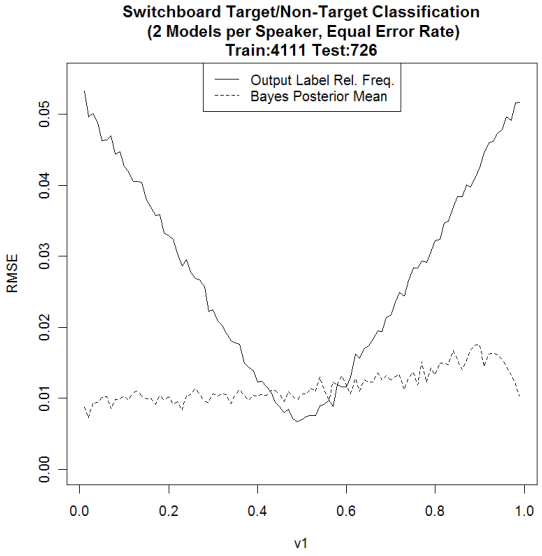

Fig. 2. Operating point: $c_{12} = c_{21} = 0.053$

4. EXPERIMENTAL EVALUATION ON SPEAKER ID

4.1. Introduction

In this section, we experimentally evaluate the M-H algorithm on a target/non-target speaker identification (SID) task derived from the Switchboard corpus [15]. We will construct this task by randomly selecting 100 speakers (out of nearly 500) to constitute a modeled target set, with the remaining open-set (unmodeled) denoted as non-targets.

We evaluate the RMSE as a function of true target proportion v_1, comparing the RMSE curves for W^* and V^*. Randomized training (L) and test (Y) sets are generated for values of v_1 in the interval $[0, 1]$. Using the algorithm of the previous section, we estimate $P(V|L, Y)$ for each data set and compute the RMSE as a function of v_1. We then compare the RMSE of the hypothesized proportion (w_1^*), and of the Bayes estimated proportion (v_1^*), as functions of the unknown true v_1.

4.2. Data and Experimental Set-up

Andrews and Hernandez [16] provided SID scores for a subset of Switchboard, using the algorithm from [17]. In particular, there are 4837 different voice cuts representing 483 different speakers. To create a target set, 100 speakers were selected at random. To provide a task with non-negligible error rate, only two trained models were retained for each of the target speakers (i.e. 200 models). No individual models were retained from the open-set (non-targets). We defined a simple binary classifier with parameter T as follows. For each voice cut:

1. Find model scores $\{s_i\}$ for the target speakers (200 scores),
2. If $\max(s_i) > T$ then classify the voice-cut as "Target", else classify as "Non-target."

The resulting classification task has an equal error rate around 5%.

We estimate RMSE(v_1^*) over various values of v_1 as follows. Generate random partitions of the 5K voice cuts into training and test sets. Denote the training sets by L_i, where the number of voice cuts N_L is constant (4111). Denote the test sets by Y_i, where the number of voice cuts N_Y is also constant (726).

The value of 'true target proportion' v_{i1} is controlled by constrained generation of the Y_i. Denote by X_{i1} the number of true target cuts in the data set Y_i, where the true proportion of target speakers in that data set is given by $v_{i1} = X_{i1}/N_Y$. Denote by $w_{i1} = Y_{i1}/N_Y$ the hypothesized target proportion in Y_i.

For each partition we estimate $P(V|Y, L)$ via M-H. Denote by N_{v1} the number of partitions (L_i, Y_i) with common v_1 (100 in our experiments). The true v_{i1} is known for each partition. This provides an empirical measure for the RMSE of an estimator at fixed true target proportion: $\text{RMSE}(v_1^*|v_1) = \left(\sum_{vi1=v1}(v_{i1}^* - v_1)^2/N_{v1}\right)^{1/2}$ and similarly for $\text{RMSE}(w_1^*|v_1)$.

We present $\text{RMSE}(v_1^*|v_1)$ and $\text{RMSE}(w_1^*|v_1)$, based upon 100 random partitions generated for every (approximate) percentile value of v_1. Figure 2 shows the two curves at the EER operating point $c_{12} = c_{21} = 0.05$. Observe that $\text{RMSE}(w_1^*)$ is quite unpredictable, ranging between 0.01 and 0.05 depending on the true value of v_1. $\text{RMSE}(v_1^*)$ is both significantly lower and more predictable. The maximum of $\text{RMSE}(v_1^*)$ is a factor of 5 smaller than the maximum of $\text{RMSE}(w_1^*)$. Furthermore, measuring the predictability of the errors by range, then $0.007 < \text{RMSE}(v_1^*) < 0.017$ at the equal operating point, while $0.006 < \text{RMSE}(w_1^*) < 0.053$. This gives a range of 0.01 versus 0.047, or a 75% relative reduction in the range of v_1^*. Figure 3 shows the estimator RMSE curves when the false alarm rate (c_{12}) is 2% and the missed detection rate (c_{21}) is 9%.

4.3. Value Estimation on Streams

One important problem is to identify which of several data streams has the greatest proportion of some target class. If all streams are have the same classifier error, the best source of the target class is the one with the highest observed w_1^*. In practice however streams often differ, for example due to noise and channel effects. In these cases hypothesized classes W alone can lead to consistently poor decisions.

We generate two data sets with different non-target distributions

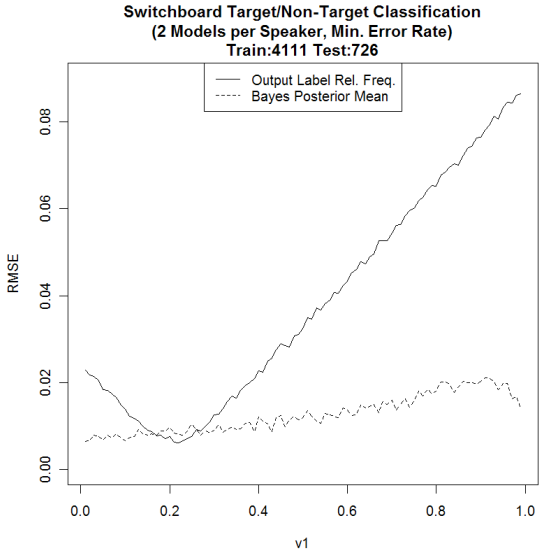

Fig. 3. Operating point: $c_{12} = 0.024$, $c_{21} = 0.088$

via biased sampling. Rather than random allocation, we assign a fixed proportion of those points on which the classifier fails to subsets of the data. In particular, we divide the Switchboard data into two halves S_1 and S_2, but allocate exactly one-third of classifier errors to S_1. This creates overall error rates of 3.6% and 7.1% on S_1 and S_2 respectively.

We partition each S_i into equal pieces L_i and Y_i. Classifier performance C is modeled independently on each L_i to allow for changes. We examine the results of estimation on 1000 partitions (L_1, L_2, Y_1, Y_2) for various fixed values of true target proportion v_1 on Y_1 and Y_2.

If we set target proportions $v_1 = 0.03$ in Y_1 and $v_1 = 0.01$ in Y_2, the mean value of w_1^* is 0.064 in Y_1 and 0.080 in Y_2. The mean value of v_1^* is 0.031 in Y_1 and 0.015 in Y_2. Bayes estimation decides Y_1 is the richer source of target voice cuts 87.7% of the time; hypothesized classes select it only 0.6% of the time.

With target proportions $v_1 = 0.047$ in Y_1 and $v_1 = 0.01$ in Y_2, the mean value of w_1^* is 0.079 in each; true target difference exactly matches the difference in bias. Means for v_1^* are 0.047 and 0.015 respectively. Bayes estimation selects the richer source 99.2% of the time. Hypothesized classes have essentially random performance (correct 47.2% of the time). Only as the difference in true class proportions increases beyond 3.5%, do hypothesized classes detect the difference. We see that given rare target classes, a difference in false alarm rate can overwhelm the difference in true target proportion.

5. CONCLUSIONS

This paper has addressed the problem of estimating class proportions based on the output of an automated pattern classification system, for example language, speaker or topic identification. We described an hierarchical Bayes model for the true class distribution, which allows construction of a Bayes estimator for the true class proportion.

This algorithm was experimentally evaluated on a binary SID task derived from the Switchboard corpus. This experiment demonstrated that the Bayes estimator of target proportion is far superior to the hypothesized target proportion from the classifier. The maximum RMSE was reduced by a factor of 5, and the range in RMSE (as a measure of variability) is reduced by a factor of 4.

6. REFERENCES

[1] A.L. Gorin, "Coping with Information Overload," in *Proceedings of the International Symposium on Large-scale Knowledge Resources*, 2006.

[2] J. Langford, "Tutorial on Practical Prediction Theory for Classification," *Journal of Machine Learning Research*, vol. 6, pp. 273–306, 2005.

[3] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, vol. 49, pp. 327–335, 1995.

[4] J. Grothendieck, "Tracking Changes in Language," *IEEE Transactions on Speech and Audio Processing*, pp. 700–711, 2005.

[5] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[6] S. Singh, C. Estan, G. Varghese, and S. Savage, "Automated Worm Fingerprinting," in *Proc. OSDI*, May 2004, pp. 45–60.

[7] Stephen G. Eick, John W. Lockwood, Ron Loui, Andrew Levine, Justin Mauger, Doyle J. Weushar, Alan Ratner, and John Byrnes, "Hardware Accelerated Algorithms for Semantic Processing of Document Streams," in *IEEE Aerospace Conference*, 2006, Paper 10.0802.

[8] N.J. Wald, K. Nanchahal, S.G. Thompson, and H.S. Cuckle, "Does Breathing Other People's Tobacco Smoke Cause Lung Cancer?," *British Medical Journal*, vol. 293, pp. 1217–1222, 1986.

[9] S.D. Walter and L.M. Irwig, "Estimation of Test Error Rates, Disease Prevalence, and Relative Risk from Misclassified Data: A Review," *Journal of Clinical Epidemiology*, vol. 41, pp. 923–937, 1988.

[10] L. Joseph, T. Gyorkos, and L. Coupal, "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard," *American Journal of Epidemiology*, vol. 141, pp. 263–272, 1995.

[11] A. Gelman, G.O. Roberts, and W.R. Gilks, "Efficient Metropolis Jumping Rules," *Bayesian Statistics 5*, pp. 599–607, 1994.

[12] H. Haario, E. Saksman, and J. Tamminen, "An Adaptive Metropolis Algorithm," *Bernoulli*, vol. 7, no. 3, pp. 223–242, 2001.

[13] E.L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, 1998.

[14] L. Tierney, "Markov Chains for Exploring Posterior Distributions (with discussion)," *Annals of Statistics*, vol. 22, pp. 1701–1762, 1994.

[15] J. Godfrey and E. Holliman, "SWITCHBOARD-1 Release 2," Linguistic Data Consortium, LDC97S62, 1997.

[16] W. Andrews and J. Hernandez-Cordero, "SREC'05 output on Switchboard I, private communication," 2006.

[17] D. Reynolds, W. Campbell, W. Shen, P. Torres-Carasquillo, and A. Adami, "MIT-Lincoln Laboratory System Description NIST SRE 2005," 2005.