

Coping with Training Contamination in Unsupervised Distributional Anomaly Detection

Nash Borges and Gerard G. L. Meyer
Human Language Technology Center of Excellence
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD 21218
Email: {nashborges,gglmeyer}@jhu.edu

Abstract—In previous work [1], we presented several distributional approaches to anomaly detection for a speech activity detector by training a model on purely nominal data and estimating the divergence between it and other input. Here, we reformulate the problem in an unsupervised framework and allow for anomalous contamination of the training data. After noting the instability of Gaussian mixture models (GMMs) in this context, we focus on non-parametric methods using regularly binned histograms. While the performance of the log likelihood baseline suffered as the amount of contamination was increased, many of the distributional approaches were not affected. We found that the L_1 distance, χ^2 statistic, and information theory divergences consistently outperformed the other methods for a variety of contamination levels and test segment lengths.

I. INTRODUCTION

Large corpora of labeled speech exist for training classifiers such as those used for speaker identification, language identification, and speech-to-text systems. While these data sets were carefully constructed to simulate real-world problems, the data is fairly well-behaved and culled of any gross anomalies deemed irrelevant to the task at hand. For such tasks, state-of-the-art systems perform quite well, especially when test data is similar to the training data. Our goal was to develop an anomaly detector to supplement these existing classifiers by recognizing data that should not be processed normally.

Supervised machine learning algorithms predict an output $y \in \mathcal{Y}$ for each input $x \in \mathcal{X}$ using a set of manually labeled training data $\{(x_1, y_1), \dots, (x_n, y_n)\}$. In a binary classification context, we can define $\mathcal{Y} = \{0, 1\}$ and use examples from both classes to develop a generative model $p(x, y)$ or a discriminative classifier $p(y|x)$.

If we label each nominal example with $y = 0$ and each anomaly with $y = 1$, anomaly detection falls naturally into this binary classification framework. Research on detecting anomalies has been conducted in many different fields, such as network intrusion detection [2], fraud detection [3], motion segmentation [4], and stylistic inconsistency detection [5]. In Hodge and Austin’s survey of the subject [6], they categorized the research into methods that required labeled examples of both classes, those that learned from only nominal data, and those that were completely unsupervised. A subset of the unsupervised methods used a strategy of *accommodation*, defined as the ability to “cope with a sizable fraction of contamination” [7].

In this work, we sought to adapt our previous approaches that required purely nominal data to methods of robust accommodation. While classification noise was studied [8] in the probably approximately correct (PAC) learning framework [9], we explored its effect when performing a detection task and learning from noisy negative examples.

A similar problem of learning from positive and unlabeled examples was first proven possible [10] in the context of the PAC framework. Here, in addition to the positive examples, the learner can also request unlabeled examples that may come from either class. In this context, Zhu [11] distinguished methods that built classifiers from those that ranked input according to their similarity to a positive query. Using an estimate of the probability of a positive example in the unlabeled set, binary classification has been performed using a decision tree [12] and naïve Bayes model [13].

Others in the text field have approached the problem differently, by first automatically labeling some “reliable” negative examples from the unlabeled set and then iteratively applying either naïve Bayes [14], an SVM [15], or logistic regression [16] to build a classifier. One of many assumptions made here is that the unlabeled set is comprised mostly of the opposite class of examples from the labeled set. When retrieving relevant documents from a set that is mostly irrelevant such an assumption is valid, but the opposite is usually true for anomaly detection where the probability of occurrence is consistently small. Our approach is more akin to the ranking procedure described by Zhu [11], where we wish to assign a measure of anomalousness to each audio segment relative to data that is *mostly* nominal.

II. DATA

A. Syllable Rate Features

One challenge in speech processing is dealing with a large quantity of data. While many speech tasks use spectral features computed every 10 ms resulting in approximately 50 dimensions, we used two features from a syllable rate speech activity detector (SRSAD) [17] computed every 100 ms. Since speech has a syllable rate of approximately 5 Hz, the frequency of its envelope modulation is different from white noise. Using a sliding half second window of audio, SRSAD computes the expected value of this modulation frequency and an estimate of

its power. We modeled the distribution of this two-dimensional sequence as a set of independent observations.

B. Anomalous Data

We developed a set of synthetic anomalies which are known to be problematic to SRSAD, such as tones and noises of short duration and certain kinds of muzak [18]. The set of anomalies was comprised of 50 examples of each of the following: DTMF sequences, morse code, MIDI tones, MIDI songs, and various telephony noises. Audacity [19] plug-in effects were used to generate 5 minute long random DTMF and morse code sequences using varying tone lengths from 25 ms to 1.25 s along with a set of always-on MIDI tones centered at frequencies from 10 Hz to 300 Hz. The MIDI songs were downloaded from the MIDI Database [20] and have a median length of 3.5 minutes. Telephony noises were obtained from FindSounds [21] using the following search terms: *busy signal, cell phone, dial tone, fax, keyboard, modem, off-hook, phone, printer, ringing, and typing*. Since some of these noises were of short duration, the audio was repeated until each was at least 5 minutes long. Half of the examples of each anomaly type were randomly selected for testing and the other half were reserved for possible use as training contamination.

C. Nominal Data

The CallHome English corpus [22] of unscripted conversational speech between family and friends was selected to represent nominal audio. Each conversation side in the *train* set was divided into 5 minute segments and 250 of the total 918 were randomly selected for training. This enabled us to experiment with contamination percentages up to 33% when using all 125 anomalous segments. The English *eval* set was similarly divided into 5 minute segments yielding 226 for testing. As our algorithms made no use of any labels, we were not restricted to the subset of audio with associated transcripts.

When testing on less than 5 minutes of audio, we randomly selected one continuous section of the desired length from each 5 minute segment. We did not investigate varying the amount of data used for training.

III. GAUSSIAN MIXTURE MODELING

We began by subjecting our previous GMM-based methods [1] to varying amounts of training contamination (Figure 1). When there was no training contamination, the approximation to the Kullback-Leibler (KL) divergence performed well for 8 and 16 GMMs. For contamination levels above 3.5% it became extremely unstable, unpredictably switching between detecting anomalies and nominal segments. While it might have been possible to devise a strategy to change the output labels if we estimated the contamination level, we chose to look elsewhere. The remainder of this paper investigates methods using non-parametric histogram-based density estimation.

IV. HISTOGRAM MODELING

Since the syllable rate features were reasonably bounded in \mathbb{R}^2 , we used histograms to model their distribution. This non-parametric density estimate is more robust to contamination

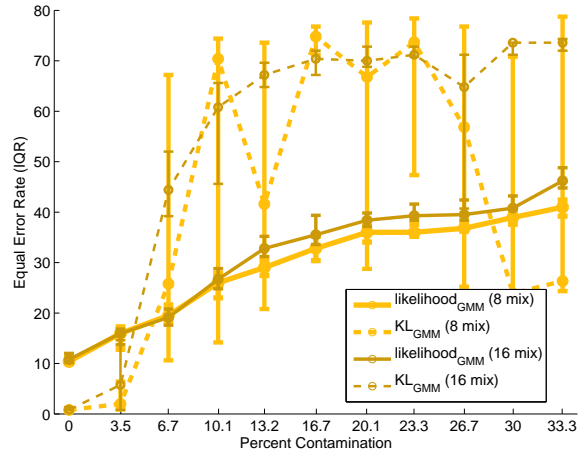


Fig. 1. Interquartile range of anomaly detection error rate for 10 GMM initializations when testing on the full 5 minute segments with varying amounts of training data contamination.

since it does not assume that the data was generated by a mixture of Gaussians. Histograms can use adaptive bins whose locations change for each input sequence or they can use fixed bins for all input sequences. While adaptive binning can result in a lower error between the feature vectors and bin centroids, we chose not to use it because it eliminates many computationally efficient histogram dissimilarity measures [23].

Techniques using fixed bins come in two varieties. Either the feature space is divided up into regular intervals or the bin locations are derived from clustering the data. The latter has been used extensively in image retrieval applications where a fixed database of images is being searched [24]. We used regular intervals to characterize the mostly nominal training data since we could not guarantee that anomalies were present when the bin locations were derived. Our main goal in unsupervised anomaly detection was recognizing unexpected anomalies long after training was completed.

We define a histogram H as a set $\{(b_1, c_1), \dots, (b_m, c_m)\}$ of bin centroids b_j with corresponding counts

$$c_j = \sum_{i=1}^n I_{\{j\}} \left(\arg \min_k d(b_k, x_i) \right) \quad (1)$$

using the Euclidean distance d and set indicator function I on training data x_i . Figure 2 shows examples of this for the training data with (a) 0% contamination and (d) 33% contamination. Histograms of nominal test segments are shown in (b) and (e) alongside anomalous segments (c) and (f).

If the dissimilarity between histograms for the mostly nominal model (MNM) and a test sequence exceeded a threshold λ , the test sequence was labeled as anomalous. The threshold λ was chosen so the false alarm rate was equal to the miss rate on the test set. This equal error rate (EER) allowed us to summarize detection performance with a single number, however an *a priori* threshold [25] would be required in a non-experimental setting.

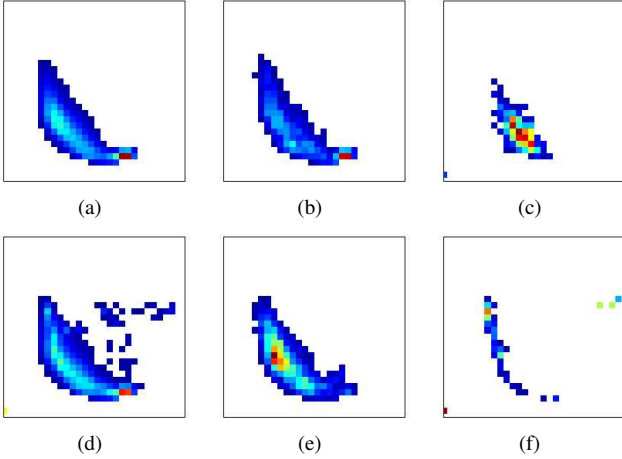


Fig. 2. Histograms of (a) 0% contaminated training data, (d) 33% contaminated training data, (b) 5 minutes of CallHome English 5888 side 2, (e) 5 minutes of CallHome English 6825 side 2, (c) MIDI song, and (f) Morse code. Using the approximation to the KL divergence, those four test segments have histogram dissimilarity measures of 0.13, 0.45, 5.83, and 9.89, respectively.

A. Log Likelihood Baseline

We first present a baseline anomaly detector using the average log likelihood of a test sequence. In this case, a test segment $X = \{x_{n+1}, \dots, x_{n+t}\}$ was labeled as anomalous if

$$\frac{1}{t} \sum_{i=n+1}^{n+t} \log p_{\text{MNM}}(x_i) < \lambda, \quad (2)$$

where

$$p_{\text{MNM}}(x_i) = \frac{c_j + \frac{1}{m}}{(\sum_{k=1}^m (c_k)) + 1} \quad (3)$$

and $j = \arg \min_k d(b_k, x_i)$ for the MNM histogram with m cells. The smoothed cell probability estimate in Equation 3 was also used in other dissimilarity measures when required.

While estimating the likelihood of individual feature vectors allows decisions to be made on a much shorter time scale, we believed that the distribution of features over time would provide better evidence for deciding between nominal and anomalous audio. We were also interested in determining both if the histogram-based KL divergence with which we had previous success [1] would be robust to training contamination and how well it would compare to other histogram dissimilarity measures described by Rubner et al. [24]

B. Minkowski-form Metrics

To compare two histograms, H_1 as the MNM and H_2 a test segment, we began by computing their L_1 , L_2 , and L_∞ distances using the following,

$$D_{L_r}(H_1, H_2) = \left(\sum_{j=1}^m |p_{H_1}(b_j) - p_{H_2}(b_j)|^r \right)^{\frac{1}{r}}. \quad (4)$$

The L_1 distance is the sum of the absolute cell differences and has been used to compute color dissimilarity between images [26]. L_2 is the sum of the squared differences and L_∞

is the max difference that has been used to compute texture dissimilarities [27].

C. Test Statistics

We used three different statistics to test the null hypothesis that the test segment was generated from the same probability distribution as the MNM.

1) χ^2 statistic: We modified Puzicha et al.'s proposal [28] for histogram-based image retrieval,

$$D_{\chi^2}(H_1, H_2) = \sum_{j=1}^m \frac{(p_{H_2}(b_j) - p_{H_1}(b_j))^2}{p_{H_1}(b_j)} \quad (5)$$

to test if H_2 differed from the mostly nominal H_1 .

2) *Kolmogorov-Smirnov statistic*: Defined as the maximal difference between one-dimensional empirical cumulative distribution functions, a histogram-based approximation

$$D_{\text{KS}}(H_1, H_2) = \max_j |P_{H_1}(b_j) - P_{H_2}(b_j)|, \quad (6)$$

was proposed by Geman et al. [29] for grayscale boundary detection. The marginal distributions are often used in a multidimensional setting, but we took a different approach. The cells from the two dimensional histograms were ordered by descending count of the MNM and cumulatively summed to obtain each P .

3) *Cramér-von Mises criterion*: Using the same strategy of converting the two-dimensional histograms to a one-dimensional cumulative density function, we approximated this statistic with

$$D_{\text{CvM}}(H_1, H_2) = \sum_{j=1}^m (P_{H_1}(b_j) - P_{H_2}(b_j))^2. \quad (7)$$

D. Information Theory Divergences

Shannon formalized information theory as a study of communication and channel capacity in the presence of noise [30]. Kullback and Leibler followed by generalizing the concept of information in their study of the “statistical problem of discrimination” between distributions [31], the results of which we were interested in using for anomaly detection.

1) *Kullback-Leibler divergence*: The mean information for discrimination

$$D_{\text{KL}}(H_1, H_2) = \sum_{j=1}^m p_{H_1}(b_j) \log \frac{p_{H_1}(b_j)}{p_{H_2}(b_j)}, \quad (8)$$

is essentially a measure of how well one statistical population can be compressed using the other as a codebook [32].

2) *Jensen-Shannon divergence*: A symmetrization of the KL divergence proposed by Lin [33], we also computed

$$D_{\text{JS}}(H_1, H_2) = \frac{1}{2} D_{\text{KL}}(H_1, G) + \frac{1}{2} D_{\text{KL}}(H_2, G) \quad (9)$$

where $p_G(x) = \frac{1}{2} p_{H_1}(x) + \frac{1}{2} p_{H_2}(x)$, which added numerical stability and bounded the divergence between 0 and 1.

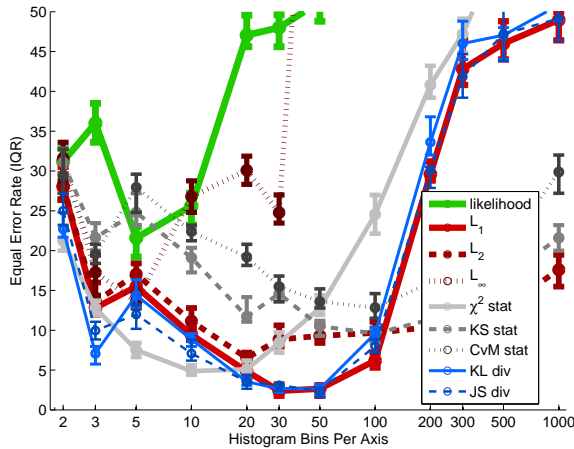


Fig. 3. Interquartile range of anomaly detection equal error rate as a function of model complexity with 33% anomalous contamination of the training data for 30 second test segments.

V. RESULTS

A. Model Selection

Model complexity had a substantial impact on anomaly detection performance, especially for short test segments. In Figure 3, we show the performance of each approach on 30 second segments as the number of histogram bins was varied when training with 33% anomalous contamination. The interquartile range of the EER was derived from random resamplings of the test set via statistical bootstrapping [34]. The log likelihood and L_∞ distance made their fewest errors when using only 5 bins per axis, the χ^2 statistic when using 10 bins per axis, and the L_2 distance when using 20 bins per axis. The L_1 distance and information theory divergences achieved their optimum performance with 30 bins per axis, and the Kolmogorov-Smirnov (KS) test and Cramér-von Mises (CvM) criterion did slightly better using 50 bins per axis. These histogram bin sizes were used for the remainder of the results as we varied other parameters.

This sensitivity is similar to the bias-variance trade-off in parameter estimation and optimizing for it is important when engineering statistical systems. When the model complexity was too low, we could accurately estimate the cell probabilities, but the restricted model space was “biased” away from the true distribution and performance suffered. As the bin sizes decreased, the model space expanded yielding the ability to better model the true distribution. However, when the bins became too small the increased variance in probability estimates eclipsed modeling power, again leading to a decrease in performance.

When we tested on the entirety of the 5 minute segments (Figure 4), the models performed well for a much wider range of complexities. With more data we could accurately estimate more parameters, which lead to better performance when using higher complexity models.

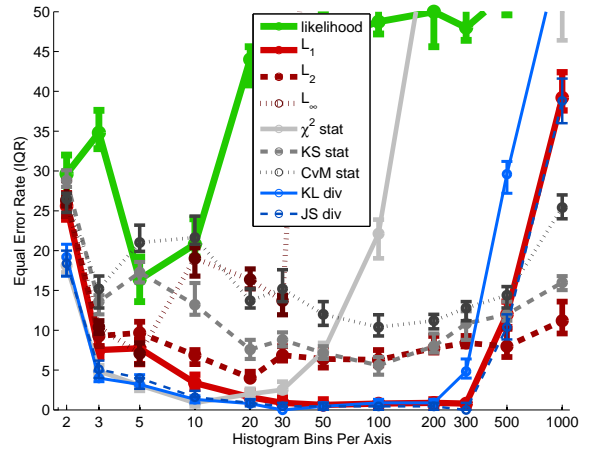


Fig. 4. Interquartile range of anomaly detection equal error rate as a function of model complexity with 33% anomalous contamination of the training data for the full 5 minute test segments.

B. Test Segment Length

To further explore the effect of test segment length, we evaluated anomaly detection performance when training with 33% contamination as we varied the length of the test segments from 5 seconds to 5 minutes (Figure 5). Some general rankings of the various approaches began to emerge, especially for tests using longer segments. The L_1 distance, χ^2 statistic, and both information theory divergences significantly outperformed the other methods when testing on segments longer than 15 seconds. The KL divergence made *no errors* on the full 5 minute segments and the JS divergence, χ^2 statistic, and L_1 distance achieved EERs of 0.8%, 0.9%, and 1.3%, respectively. These four methods were also among the best performers on the shorter segments, but the differences were not always significant.

The L_2 distance was the next best performer with an EER of 4.4% on the 5 minute segments. The KS test and L_∞ distance had comparable EERs of 7.2% and 8% while the CvM criterion and log likelihood baseline did not perform as well with EERs of 12.4% and 16.8%, respectively.

To explore the interplay between test segment length and model complexity, we show the EER for a range of both (Figure 6) when using the L_1 distance, χ^2 statistic, and KL divergence. We see that the L_1 distance is especially attractive because its simplicity and numerical stability lead to a robustness unparalleled by the other two approaches. We see again that the χ^2 statistic performed well with 10 bins per axis while the KL divergence was harder to predict. We report the performance of all the approaches when testing on several segment lengths in Table I.

C. Training Data Contamination

The aim of this work was to relax our previous assumption that we had a large amount of *purely* nominal data for training the MNM. It is easier and far less expensive to obtain mostly nominal, unlabeled data, as it does not require any human labeling effort. Using unlabeled data also allowed us

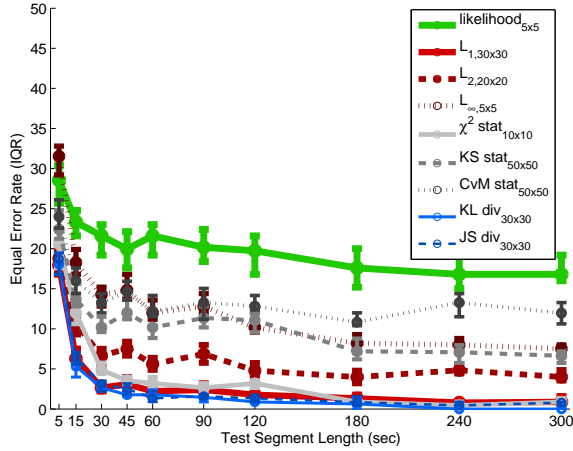


Fig. 5. Interquartile range of anomaly detection equal error rate as a function of test segment length for 33% anomalous training contamination.

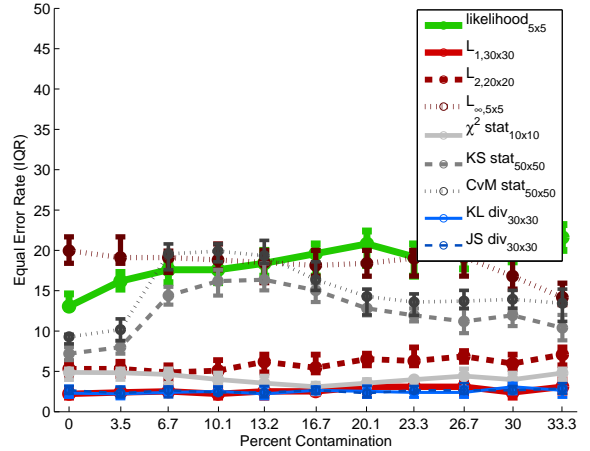


Fig. 7. Interquartile range of anomaly detection equal error rate as a function of the amount of training data contamination for 30 second test segments.

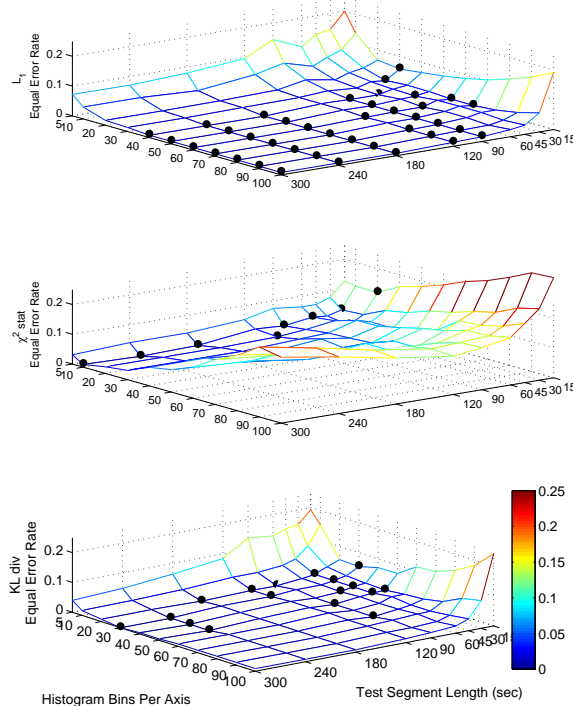


Fig. 6. Anomaly detection equal error rate as the model complexity and test segment length are varied when training for 33% anomalous training contamination. A black dot is placed at the minimum(s) of each test segment length.

to avoid the quagmire of having to define what constitutes an anomalous sound. We preferred to keep the concept of an anomaly somewhat vague since we were attempting to find audio that is anomalous from the perspective of speech processing algorithms, not humans.

We therefore investigated if our methods could robustly model the nominal data even if it was partially contaminated with anomalies. This was tested by incrementally adding anomalies into the data used for training the MNM and evaluating the anomaly detection performance. At each con-

tamination level from 0% to 33% in approximate increments of $3\frac{1}{3}\%$, we display the interquartile range of the EER via statistical bootstrapping (Figure 7).

After seeing the effect of contamination on the GMM-based methods (Figure 1), we were encouraged by the robustness of all of the histogram-based approaches. The information theory divergences, χ^2 statistic, and L_1 and L_2 distances were minimally affected by the contamination. Our strategy of cell-ordering based on the MNM for the KS test and CvM criterion did not fare as well. With purely nominal data, the EER was cut in half as compared to an arbitrary ordering, but the effect quickly diminished as the contamination level reached 6.7%. For higher contamination levels performance started to improve again. We attribute this unexpected behavior to our optimization of model complexity at 33% contamination. This also explains the performance of the L_∞ distance that was inversely related to the contamination level. The log likelihood baseline was the only method negatively affected by the amount of contamination. The slight effect seen here was much less pronounced than the effect seen for slightly more complex models.

We end our analysis of these methods by evaluating the detection error trade-off curve [35] when training at 33% contamination and testing on 30 second segments (Figure 8). The L_1 distance and information theory divergences were the best performers in this harsh test condition achieving EERs of 2.7% and 3.1%, respectively. The χ^2 and L_2 distance formed the next tier of performers with EERs of 4.9% and 6.4%, followed by KS, CvM, and L_∞ with EERs between 10.2% and 13.7%. The log likelihood baseline remained the worst performer with a 20.8% EER.

VI. CONCLUSION

While the KL divergence approximation for GMMs [1] performed well when the MNM was trained with purely nominal data, we found that performance quickly degraded with more than 3.5% contamination. In comparison, the histogram-based methods presented here were barely affected by up to

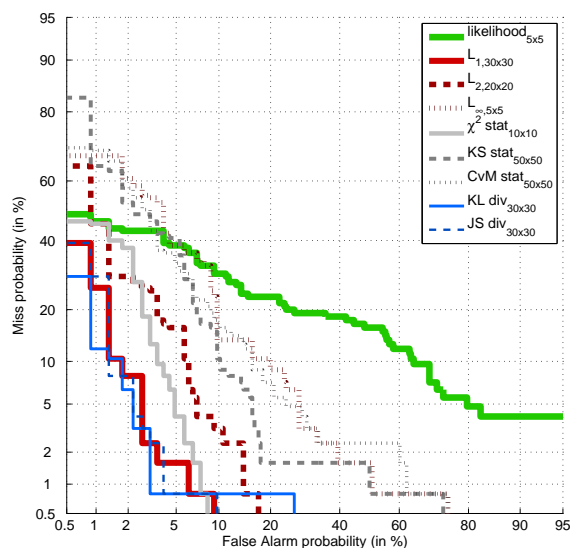


Fig. 8. Anomaly detection error trade-off curve when training with 33% contaminated data and testing on 30 second segments.

TABLE I
PERCENT EQUAL ERROR RATE.

Test Length (sec)		5	15	30	60	300
0% Contamination	likelihood _{5x5}	21.6	15.2	14.4	12.0	10.2
	L _{1,30x30}	13.3	4.9	2.4	1.3	0.4
	L _{2,20x20}	14.4	8.8	4.9	4.0	2.7
	L _{∞,5x5}	36.0	30.1	19.5	16.8	12.0
	χ ² stat _{10x10}	12.0	8.4	4.9	4.4	3.5
	KS stat _{50x50}	17.6	10.4	7.2	7.2	4.8
	CvM stat _{50x50}	18.1	12.8	9.6	8.0	5.8
	KL div _{30x30}	11.5	4.0	2.7	0.9	0.0
JS div _{30x30}	12.4	4.9	2.7	1.3	0.4	
10% Contamination	likelihood _{5x5}	22.4	17.6	17.6	14.6	13.6
	L _{1,30x30}	15.2	4.9	2.7	1.6	0.4
	L _{2,20x20}	15.0	8.4	7.2	6.2	3.1
	L _{∞,5x5}	30.1	29.6	18.6	16.4	13.6
	χ ² stat _{10x10}	13.3	8.4	3.5	3.1	1.6
	KS stat _{50x50}	25.6	17.6	16.8	11.2	11.2
	CvM stat _{50x50}	28.0	21.2	20.8	17.3	14.2
	KL div _{30x30}	12.8	4.0	2.7	1.3	0.0
JS div _{30x30}	13.3	4.9	2.7	1.3	0.4	
33% Contamination	likelihood _{5x5}	27.2	23.5	22.1	20.8	16.8
	L _{1,30x30}	17.7	5.8	2.7	1.8	1.3
	L _{2,20x20}	18.1	10.6	6.4	5.8	4.4
	L _{∞,5x5}	31.4	18.6	13.6	11.9	8.0
	χ ² stat _{10x10}	20.8	12.0	4.9	4.0	0.9
	KS stat _{50x50}	21.7	13.3	10.2	8.8	7.2
	CvM stat _{50x50}	24.0	16.4	13.7	13.6	12.4
	KL div _{30x30}	18.1	5.3	3.1	1.6	0.0
JS div _{30x30}	18.6	6.6	3.1	1.8	0.8	

33% training contamination. The L_1 distance, χ^2 statistic, and information theory divergences typically were the best performers. While the KL divergence made no errors on the full 5 minute test segments regardless of the contamination level, we prefer the L_1 distance since its performance was comparable and its simplicity yielded the most robust performance to model complexity and test segment length.

REFERENCES

- [1] N. Borges and G. G. L. Meyer, "Unsupervised distributional anomaly detection for a self-diagnostic speech activity detector," in *CISS*, 2008.
- [2] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *SIAM International Conference on Data Mining*, 2003.
- [3] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," in *Credit Scoring Conference*, 2001.
- [4] P. Torr and D. Murray, "Outlier detection and motion segmentation," in *Proc. of SPIE*, vol. 2059, 1993, pp. 432–443.
- [5] D. Guthrie, L. Guthrie, B. Allison, and Y. Wilks, "Unsupervised anomaly detection," in *IJCAI*, 2007.
- [6] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.
- [7] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Wiley, 1987.
- [8] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, no. 4, pp. 343–370, 1988.
- [9] L. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.
- [10] F. Denis, "PAC learning from positive statistical queries," in *Algorithmic Learning Theory*, 1998, pp. 112–126.
- [11] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2007.
- [12] F. Letouzey, F. Denis, and R. Gilleron, "Learning from positive and unlabeled examples," in *Algorithmic Learning Theory*, 2000, pp. 71–85.
- [13] F. Denis, R. Gilleron, and M. Tommasi, "Text classification from positive and unlabeled examples," in *IPMU*, 2002.
- [14] B. Liu, W. Lee, P. Yu, and X. Li, "Partially supervised classification of text documents," in *ICML*, 2002, pp. 387–394.
- [15] H. Yu, J. Han, and K. Chang, "PEBL: positive example based learning for web page classification using SVM," in *KDD*, 2002.
- [16] W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," in *ICML*, 2003.
- [17] D. J. Nelson, D. C. Smith, and J. L. Townsend, "Voice activity detector," US Patent No. 6556967, April 2003.
- [18] D. C. Smith, J. Townsend, D. J. Nelson, and D. Richman, "A multivariate speech activity detector based on the syllable rate," in *ICASSP*, 1999.
- [19] "Audacity: Plug-In Effects," 2009. [Online]. Available: <http://audacity.sourceforge.net/download/plugins>
- [20] "MIDI Database," 2009. [Online]. Available: <http://www.mididb.com>
- [21] "FindSounds," 2009. [Online]. Available: <http://www.findsounds.com>
- [22] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech," Linguistic Data Consortium, Philadelphia, 1997.
- [23] W. Leow and R. Li, "The analysis and applications of adaptive-binning color histograms," *CVIU*, vol. 94, no. 1-3, pp. 67–91, 2004.
- [24] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *CVIU*, vol. 84, no. 1, pp. 25–43, 2001.
- [25] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *ICASSP*, 1988.
- [26] M. Swain and D. Ballard, "Color indexing," *IJCV*, vol. 7, no. 1, pp. 11–32, 1991.
- [27] H. Voorhees and T. Poggio, "Computing texture boundaries from images," *Nature*, vol. 333, no. 6171, pp. 364–367, 1988.
- [28] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *CVPR*, 1997.
- [29] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *PAMI*, vol. 12, no. 7, pp. 609–628, 1990.
- [30] C. Shannon, "A mathematical theory of communications," *Bell System Technical Journal*, vol. 27, no. 2, pp. 632–656, 1948.
- [31] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [32] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth mover's distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.
- [33] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, 1991.
- [34] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Mathematical Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
- [35] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Eurospeech*, 1997.