To appear, *IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2013.*

Distribution-Sensitive Learning for Imbalanced Datasets

Yale Song[1], Louis-Philippe Morency[2], and Randall Davis[1]
[1] MIT Computer Science and Artificial Intelligence Laboratory
[2] USC Institute for Creative Technology
{yalesong,davis}@csail.mit.edu, morency@ict.usc.edu

***Abstract*— Many real-world face and gesture datasets are by nature imbalanced across classes. Conventional statistical learning models (e.g., SVM, HMM, CRF), however, are sensitive to imbalanced datasets. In this paper we show how an imbalanced dataset affects the performance of a standard learning algorithm, and propose a distribution-sensitive prior to deal with the imbalanced data problem. This prior analyzes the training dataset before learning a model, and puts more weight on the samples from underrepresented classes, allowing all samples in the dataset to have a balanced impact in the learning process. We report on two empirical studies regarding learning with imbalanced data, using two publicly available recent gesture datasets, the Microsoft Research Cambridge-12 (MSRC-12) and NATOPS aircraft handling signals datasets. Experimental results show that learning from balanced data is important, and that the distribution-sensitive prior improves performance with imbalanced datasets.**

I. INTRODUCTION

Collecting a dataset of human behaviors, e.g., facial expressions and body gestures, is a time-consuming and expensive procedure. One of the main difficulties is balancing the class distribution, i.e., the number of samples per class. In many real-world scenarios, some samples are far more common than others. In face and gesture datasets [8, 12, 13], in particular, samples of the normal facial expression will be abundant, while samples of other expressions (e.g., pain [12] and various others [8]) will be difficult to obtain. As another example, in anomaly detection, normal patterns of pedestrian movements are common (by definition) compared to anomalous patterns (e.g., the circulation of non-pedestrian entities in the walkways or anomalous pedestrian motion patterns [13]).

While many effective statistical learning algorithms have been developed, such as decision trees [18], Neural Networks [1], Hidden Markov Models [19], Support Vector Machines [24], and Conditional Random Fields [9], the standard formulations of these models are sensitive to imbalanced data. Suppose we have highly skewed data, e.g., a 1:10000 ratio of positive and negative samples. Using standard algorithms, the learning process will be dominated by the negative class, which in turn will classify most test samples as negative [25].

One possible solution to this problem is to balance the original dataset by re-sampling, e.g., by random undersampling or oversampling [10]. However, this approach has its drawbacks, such as potentially removing important examples in undersampling, and adding redundant examples in oversampling, which may cause overfitting [14].

In this paper we propose a distribution-sensitive prior to solve this problem, analyzing the training dataset before learning a model, and putting more weight on the samples from underrepresented classes. This allows all samples in the dataset to have balanced impact in the learning process. We report on two empirical studies regarding learning with imbalanced data, using two publicly available recent gesture datasets, the Microsoft Research Cambridge-12 (MSRC-12) [6] and the Naval Air Training and Operating Procedures Standardization (NATOPS) [21] datasets. The first experiment aims to show the effect of imbalanced data on the performance of a learned model; the second experiments evaluates the use of our distribution-sensitive prior.

Section II reviews some of the previous approaches to imbalanced data learning, Section III formalizes the problem and describes our distribution-sensitive prior, and Section IV discusses the two empirical studies. Section V summarizes the findings and suggests directions for the future work.

II. RELATED WORK

There are three main groups of solutions to learning with imbalanced data: sampling methods, cost-sensitive learning, and one-class learning. In this section, for each group of solutions, we briefly introduce the core idea, point to some representative work, and highlight the differences between our method and previous work. For a comprehensive survey, readers are referred to [7].

Sampling methods use various mechanisms to modify the original dataset so that it has a balanced distribution. This is motivated by empirical studies showing that balanced data improves overall classification performance [25]. The simplest technique in this category is random undersampling, i.e., discard samples chosen randomly from overrepresented classes, or random oversampling, i.e., replicate samples chosen randomly from underrepresented classes [4]. Advanced techniques include generating synthetic data for the underrepresented class [3]. Our approach is similar to sampling methods in that, in the form of normalized weighting, the distribution-sensitive prior "simulates" duplicating samples from underrepresented class and discarding samples from overrepresented class. The difference, however, is that our approach does not require selecting which samples to duplicate/discard, nor does it require generating synthetic data.

Cost-sensitive learning takes an opposite approach by learning from imbalanced data directly, and using a misclassification cost metric that varies depending on applications

and algorithms. Elkan [5] gives a rigorous theoretical foundation for this method. Turney [23] proposed an approach that utilizes misclassification costs in the fitness function of the genetic algorithm, while Ling *et al.* [11] incorporated misclassification costs in building a decision tree. The idea of cost-sensitive learning has been applied to face recognition [26] in computer vision community. Our approach is similar to cost-sensitive learning in that minority samples are weighted higher in the learning process. The difference, however, is that we weight minority samples higher whether or not they are misclassified. This allows all samples to have balanced impacts in the learning process.

Finally, one-class learning takes the problem to an extreme and learns from the data collected from one class only. It is typically used in an outlier/anomaly detection setting. Schölkopf *et al.* [20] and Tax and Duin [22] have independently proposed a method extending a support vector algorithm to obtain a tight boundary of the one-class sample space, accepting most samples in the training dataset. Usually, kernel methods are used to obtain the boundary in a high dimensional space. We concentrate on multi-class learning rather than one-class learning, with an assumption that we have at least one sample per class.

III. DISTRIBUTION-SENSITIVE PRIOR

In this section, we formalize the problem of learning with imbalanced data, and introduce the distribution-sensitive prior to deal with this problem, which can be used in conjunction with many standard learning algorithms

A. Standard Formulation of Statistical Learning Algorithms

The heart of many statistical learning algorithms involves solving an optimization problem with an objective function $L(\mathbf{w})$ with respect to the model parameters $\mathbf{w}$. Given a training dataset of $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{R}^d, y_i \in \mathcal{Y}\}_{i=1}^N$, where x_i is a real-valued d-dimensional input feature vector and y_i is the output label, the standard way to formulate an objective is

$$\min_{\mathbf{w}} L(\mathbf{w}) = \Omega(\mathbf{w}) + Q_{emp}(\mathcal{D}, \mathbf{w}) \tag{1}$$

where $\Omega(\mathbf{w})$ is a regularizer that prevents overfitting, and $Q_{emp}(\mathcal{D}, \mathbf{w})$ is an empirical quality measure of the solution $\mathbf{w}$ derived from the training dataset $\mathcal{D}$. There exist many different definitions of both the regularizer $\Omega(\mathbf{w})$ and the empirical quality measure $Q_{emp}(\mathcal{D}, \mathbf{w})$. Typical examples of $\Omega(\mathbf{w})$ include the L_1 norm and the L_2 norm [15]. The empirical quality measure $Q_{emp}(\mathcal{D}, \mathbf{w})$ is defined differently in different learning algorithms, including the empirical risk [24], the energy [2], and the negative conditional log-likelihood probability [9, 17, 19].

B. Imbalanced Data Problem

In this paper, for ease of understanding, we concentrate on the classification problem, where $\mathcal{Y}$ is a finite alphabet set, and explain our idea using the negative conditional log probability,

$$Q_{emp}(\mathcal{D}, \mathbf{w}) = -\sum_{i=1}^{N} \log p(y_i \mid x_i) \tag{2}$$

Since we minimize $Q_{emp}(\mathcal{D}, \mathbf{w})$, this formulation maximizes a linear sum of the conditional log probabilities computed from each of the training samples (x_i, y_i) using the current solution $\mathbf{w}$.

Unfortunately, this formulation is distribution-*insensitive*: It treats each $p(y_i \mid x_i)$ as equally important and computes a linear sum of them, with an assumption that the training dataset is uniformly distributed across classes. As a result, if the dataset is highly skewed (e.g., 1:10000 ratio of positive and negative samples), the linear sum in Equation 2, and hence the solution $\mathbf{w}$, will be dominated by the most frequent classes. This will in turn classify most test samples as one of the dominating classes (as shown in [25]).

C. Distribution-Sensitive Prior

To deal with the imbalanced data problem, we define a distribution-*sensitive* prior γ_i as

$$\gamma_i = \left(\frac{\overline{N}}{N_{y_i}}\right)^k, \overline{N} = \frac{1}{|\mathcal{Y}|}\sum_y N_y \tag{3}$$

where N_y is the number of samples with a class label y (similarly for y_i), $|\mathcal{Y}|$ is the number of classes, and $\overline{N}$ is an average number of samples per class. The degree k controls the magnitude of the distribution-sensitive prior. This is then multiplied with the log probability for each sample (x_i, y_i),

$$Q_{emp}(\mathcal{D}, \mathbf{w}) = -\sum_{i=1}^{N} \gamma_i \log p(y_i \mid x_i), \tag{4}$$

When the dataset has a uniform distribution (i.e., all N_y's are the same), or when k = 0, Equation 4 is reduced to the standard formulation of Equation 2. This prior puts more weight on the samples from underrepresented classes, allowing all samples in the dataset to have balanced impact in the learning process.

In this work, we use a standard sequence classification algorithm, the Hidden Conditional Random Field (HCRF) [17], to evaluate the distribution-sensitive prior. We augment the standard formulation of HCRF with our distribution-sensitive prior γ_i:

$$\min_{\mathbf{w}} L(\mathbf{w}) = \frac{1}{2\sigma^2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \gamma_i \log p(y_i \mid x_i) \tag{5}$$

The first term in Equation 5 is the log of a Gaussian prior with variance σ^2, $p(\mathbf{w}) \sim \exp(\frac{1}{2\sigma^2}\|\mathbf{w}\|^2)$, and $p(y_i \mid x_i)$ in the second term is defined as

$$p(y_i \mid x_i) = \frac{\sum_h \exp(\mathbf{w} \cdot \Phi(y_i, h, x_i))}{\sum_{y,h} \exp(\mathbf{w} \cdot \Phi(y, h, x_i))} \tag{6}$$

where $h \in \mathcal{H}$ is a set of additional latent variables that capture the hidden dynamics in the data. The feature function

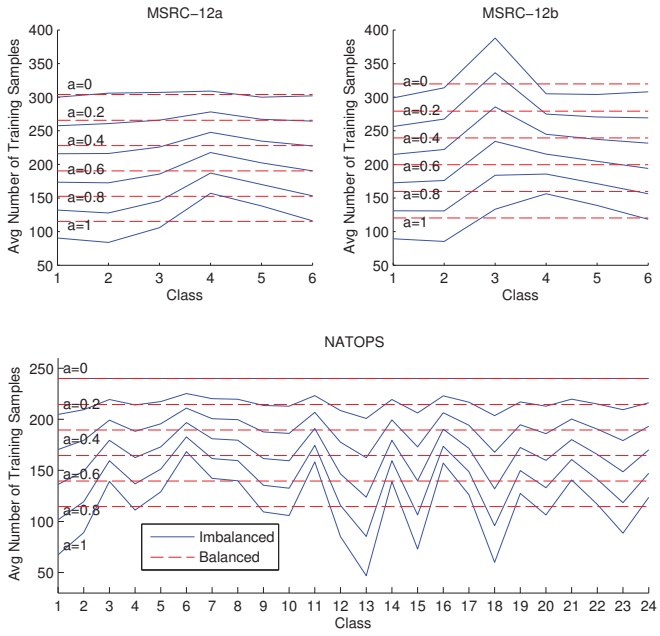

Fig. 1. Class distributions of training data with varying α. The blue lines show distributions of the imbalanced data, and the red dashed lines show the balanced version of the same data, i.e., the total amount is the same. The bias for/against particular classes is an artifact of the random seed that we set manually.

$\Phi(y, h, x)$ is derived from an underlying graph structure, which we chose a linear chain model, as is common for sequence data. Following [17], we define three types of feature functions. Let $\mathbb{1}[\cdot]$ be an indicator function, and $y' \in \mathcal{Y}$ and $h', h'' \in \mathcal{H}$. The *observation* function $\phi(h_t, x_t) = \mathbb{1}[h_t = h']x_t$ captures the compatibility between a latent variable h_t and an observation x_t; the *label* feature function $\phi(y, h_t) = \mathbb{1}[y = y' \wedge h_t = h']$ captures the compatibility between a label y and a latent variable h_t; and the edge feature function $\phi(y, h_s, h_t) = \mathbb{1}[y = y' \wedge h_s = h' \wedge h_t = h'']$ captures the compatibility among a label y and two latent variables h_s, h_t. We use the limited-memory BFGS (L-BFGS) [16] to optimize Equation 5.

IV. EXPERIMENTS

In this section, we present two empirical studies on learning with imbalanced datasets: The first experiment shows the effect of imbalanced data on the performance of a learned classifier, the second evaluates the ability of the distribution-sensitive prior to deal with imbalanced data problem. We used two publicly available recent gesture datasets, MSRC-12 [6] and NATOPS [21], selected because they are well balanced; this allows us to control the degree of imbalance manually, and see precisely how the imbalanced data affects the performance of a standard learning algorithm.

A. Datasets

MSRC-12 [6]: The Microsoft Research Cambridge-12 (MSRC-12) gesture dataset contains 12 classes of gestures divided into two categories: the iconic gesture category (crouch or hide, shoot a pistol, throw an object, change weapon, kick, and put on night vision goggles) and the metaphoric gesture category (start music/raise volume, navigate to next menu, wind up the music, take a bow to end music session, protest the music, and move up the tempo of the song). The dataset includes automatically tracked full body postures, estimated using the Kinect pose estimation pipeline. The body feature includes 3D locations of twenty body joints and is represented as a 60D feature vector. We defined each sequence as the frames in between two segment points provided in the dataset; we dropped 22 sequences whose length was longer than 400 frames. This resulted in a class distribution of [498 508 511 515 498 502] for the iconic gestures and [497 552 646 507 506 512] for the metaphoric gestures. The original data was recorded at 30 Hz; we subsampled the data by the factor of 3, resulting in 10 Hz. The average frame length of the resulting sequences was 33 for the iconic gestures and 35 for the metaphoric gestures. We normalized each dimension to have a mean zero and standard deviation one. We performed experiments on each category of gesture individually; we name these MSRC-12a (iconic gestures) and MSRC-12b (metaphoric gestures).

NATOPS [21]: The NATOPS aircraft handling signals dataset contains 24 classes of gestures used in routine practice on the deck of an aircraft carrier, e.g., turn left/right, brakes on/off, insert/remove chocks, etc. (see [21] for a complete list). The dataset includes automatically tracked upper body postures and the shapes of both hands. The body feature includes 3D velocities of four body joints – left/right elbows and wrists – and is represented as a 12D feature vector. The hand feature includes probability estimates of five predefined hand shapes – opened/closed palm, thumb up/down, and "no hand". The fifth shape, no hand, was dropped in the final representation, resulting in an 8D feature vector. The dataset has a perfectly balanced class distribution; there are 400 samples per class, obtained from 20 subjects repeating each gesture 20 times. The original data was recorded at 20 Hz; we subsampled the data by the factor of 2, resulting in 10 Hz. The average frame length of the resulting sequences was 25. We normalized each dimension to have a mean zero and standard deviation one.

B. Experiment I: Effect of Imbalanced Data

This experiment examines the effect of imbalanced data on the classification performance. Specifically, we vary the degree of imbalance by discarding samples at random with a specified degree α, and compare classifier performance to that obtained from a balanced version of the same dataset, in which the total number of training samples are the same but the class distribution is balanced across classes. This allows us to perform fair comparisons under the same amount of training samples.

1) Methodology: We performed 5-fold cross validation, where three-fifths of the entire dataset is used for training, one-fifth is used for validation, with the rest used for testing, repeated five times.

To simulate each training split having an imbalanced distribution with a degree of $\alpha \sim [0, 1]$, for each class y,

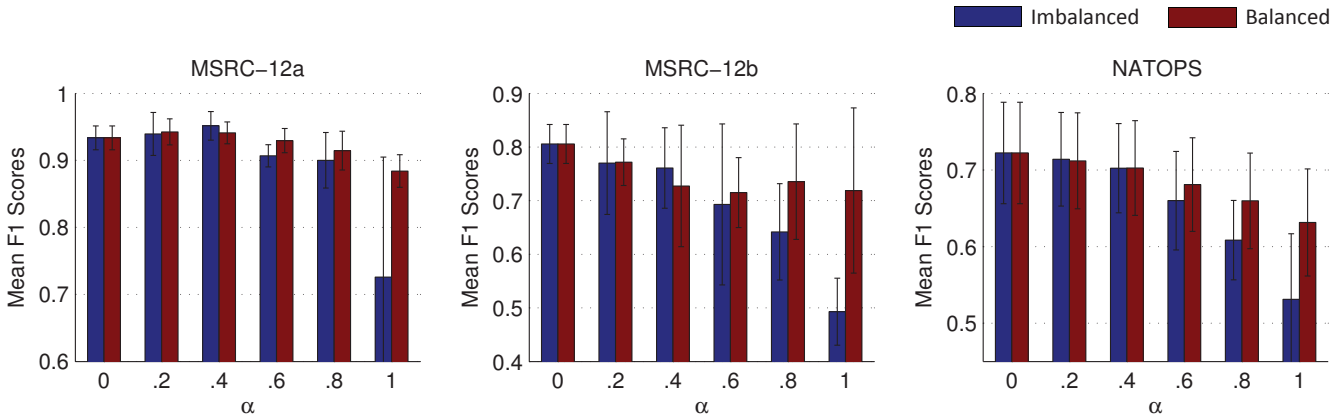

Fig. 2. The mean F1 scores as the function of α, obtained from the imbalanced and the balanced training data of the same amount. The difference of classification performance between the imbalanced and the balanced data becomes bigger as the degree of imbalanced distribution increases.

we used only the first $N'_y = N_y - rand(0, \alpha N_y)$ samples in the original training split, where N_y is the number of samples with a class label y in the original training split, and $rand(a, b)$ is a random integer number between a and b. To obtain a balanced version of the same training split, we used the first $\overline{N'} = \frac{1}{|\mathcal{Y}|}\sum_y N'_y$ samples (i.e., the average of N'_y's) from each class in the original training split. This made the total number of samples the same as the imbalanced version of the same split. While we altered the training split to control the degree of imbalance, the validation and testing splits were the same across different settings (and hence were always balanced). Figure 1 shows class distributions of training data with varying α, where we averaged out the numbers from the five splits. [1]

We varied the degree of imbalanced distribution α from 0 to 1 with an increment of 0.2 (α = 0 is same the original dataset). To allow direct comparisons between imbalanced and balanced distribution for each of the α values, we fixed all other hyper parameters of HCRF, i.e., the cardinality of the latent variables $|\mathcal{H}|$ = 8 and the L_2 regularization factor σ^2 = 10. [2] The L-BFGS optimization solver was set to terminate after 500 iterations.

2) Results and discussion: Figure 2 shows the mean F1 scores as a function of α, obtained from the imbalanced and the balanced training splits. It shows that, for the imbalanced data, the classification performance gets worse as the degree of imbalance increases (although in the case of MSRC-12a, the performance did not decrease until α = 0.4, which indicates that the training dataset may contain some redundant samples). A comparison to the corresponding balanced data shows that this decrease in performance is not due solely to the reduced amount of training data. The balanced data achieved better performance than the imbalanced data with the same amount of training data. Even though balancing the distribution reduced the number of samples for some classes, it improved the overall classification performance, indicating that having a uniformly distributed data is important.

Figure 3 shows a per-class mean accuracy plot obtained from the NATOPS dataset with α = 1. The exact difference of mean accuracies between the balanced and imbalanced data was [.5 9.8 3.3 5.0 0.3 -8.8 -0.3 1.5 12.5 11.5 -1.8 12.0 40.8 1.3 16.3 -5.8 3.3 36.0 -5.5 19.0 -5.5 11.3 20.5 4.8]. In general, the amount of the accuracy difference between the imbalanced and the balanced data roughly corresponds to the amount of the number of samples difference shown in Figure 1. Notably, the top two highest accuracy improvements (40.8 for class 13 and 36.0 for class 18) correspond to the two steepest valleys in Figure 1, which indicates that samples from the two classes in the imbalanced version of the data had very little impact during the learning process. The six classes with decreased accuracy (class 6, 7, 11, 16, 19, and 21) correspond to the hills above the red dashed line, i.e., the classes whose number of samples was reduced after the balancing was done. However, the decrease in accuracy for the six classes was minimal (an average of 4.46%) compared to the remaining classes (an average of 11.65%), resulting in overall a better performance.

C. Experiment II: Distribution-Sensitive Prior

Our second experiment evaluated the distribution-sensitive prior for learning with imbalanced data, comparing it to three baseline methods; learning with imbalanced data without using the distribution-sensitive prior (k = 0), and learning with balanced data with random undersampling and random oversampling. We studied how sensitive the classification performance is to the degree k of the distribution-sensitive prior (see Equation 3). We use the α = 1 version of the datasets from our previous experiment to simulate highly imbalanced data.

1) Methodology: We varied the degree k = [0 0.5 1 2] of our distribution-sensitive prior, where k = 0 means no distribution-sensitive prior was used. For the undersampling (and the oversampling) methods, we set the number of

[1] To make this experiment reproducible, we provide a Matlab script for generating the same data splits used in our experiments at http://people.csail.mit.edu/yalesong

[2] These parameter values were chosen based on a preliminary experiment validating $|\mathcal{H}|$ = [6 8 10] and σ^2 = [1 10 100] on the "full" version of the three datasets with α = 0. We then chose the parameter values that performed the best across five splits and three datasets.

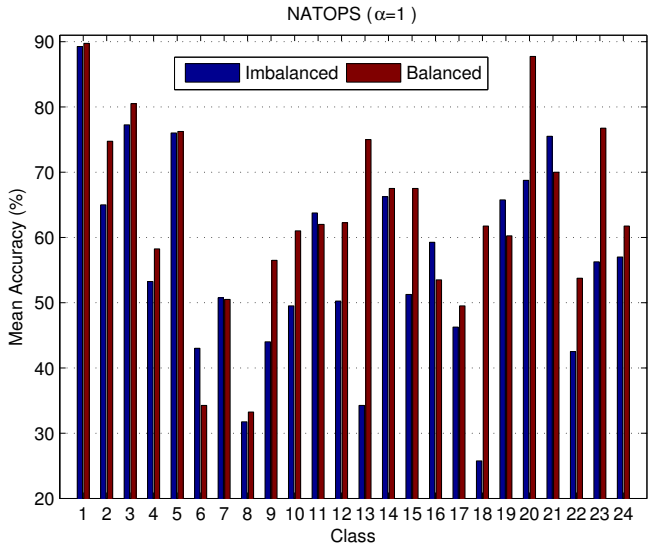

Fig. 3. Per-class mean accuracy obtained from the NATOPS dataset with α = 1, comparing the imbalanced and balanced data of the same amount. The amount of the accuracy difference between the imbalanced and the balanced data roughly corresponds to the amount of the sample count difference between the imbalanced and the balanced data shown in Figure 1.

samples per class as the minimum (and the maximum) of N'_y's, and discarded (and duplicated) samples at random to make the sample distribution even.

We validated the two hyper parameters of HCRF, the cardinality of the latent variables $|\mathcal{H}|$=[6 8 10] and the L_2 regularization factor σ^2 = [1 10 100]. We then selected, for each split and for each k, the optimal hyper parameter values based on the F1 score on the validation split. Similar to the previous experiment, we performed 5-fold cross validation, and the L-BFGS optimization solver was set to terminate after 500 iterations.

2) Results and discussion: Table I shows the mean F1 scores with standard deviations obtained from the best performing models in each method, averaged over the 5 splits. It shows that our method of using the distribution-sensitive prior outperformed the three baseline methods in all three datasets. The undersampling method performed much worse than the two others. We believe this was due to the too small amount of training samples, and that the undersampling may have discarded too many important samples.

Our method also outperformed the other baseline method, i.e., learning with imbalanced data without using a distribution-sensitive prior (k = 0). While paired t-tests between the two methods, under the same settings of $|\mathcal{H}|$ and σ^2, revealed that differences were not statistically significant (p=.55 for MSRC-12a, p=.22 for MSRC-12b, and p=.17 for NATOPS), our method did improve the performance in majority cases. Even when our method performed worse in individual test cases, the decrease in performance was minimal. For example, on the MSRC12a, 10 out of 45 cases decreased the F1 scores with an average of .016; on the MSRC-12b, 8 out of 45 cases decreased the F1 scores with an average of .022; and on the NATOPS, only 7 out of 45 cases decreased the F1 scores with an average of .004).

TABLE I

MEAN F1 SCORES WITH STANDARD DEVIATIONS FROM EXPERIMENT II.

Methods	MSRC-12a	MSRC-12b	NATOPS
Imbalanced, $k = 0$	.7808 (.14)	.5589 (.05)	.5427 (.10)
Undersampling	.6595 (.17)	.3556 (.12)	.3970 (.06)
Oversampling	.7902 (.11)	.5381 (.07)	.5779 (.08)
Dist.-sen. prior	**.7977 (.16)**	**.5721 (.06)**	**.5899 (.06)**

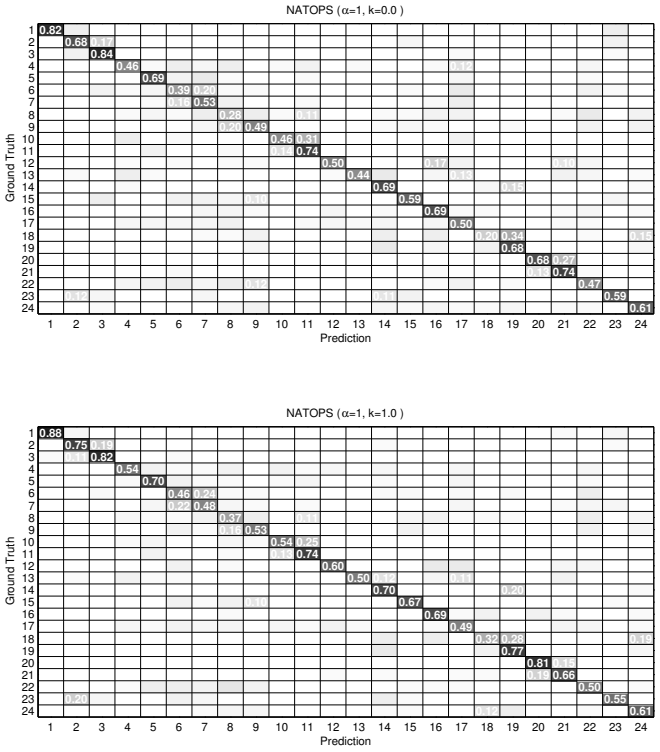

Fig. 5. Confusion matrices obtained from the NATOPS dataset at α = 1 (imbalanced), with k=0 (above; no distribution-sensitive prior used) and k=1 (bottom). Better viewed when zoomed in on a computer screen.

Figure 4 shows the mean F1 scores as a function of k, the degree of the distribution-sensitive prior. We found that setting k = 1 in general performed well, although the differences were not significant across different values of k.

Figure 5 shows two confusion matrices obtained from the NATOPS dataset at α = 1 (imbalanced), comparing k = 0 (no distribution-sensitive prior used) to k = 1. The per-class accuracy improvement was [.06 .07 -.02 .08 .02 .07 -.05 .09 .04 .08 .01 .10 .06 .01 .07 .01 -.01 .12 .09 .12 -.08 .04 -.05 0], which shows that the per-class accuracy was improved for 18 out of 24 classes. Interestingly, the accuracies of the underrepresented classes (the number of samples are lower than the average, i.e., the class below the red dashed line for α=1 in Figure 1) were all improved, except for the class 23.

V. CONCLUSIONS

In this paper, we showed how an imbalanced dataset affects the performance of a standard sequence classification algorithm using HCRF, and proposed a distribution-sensitive

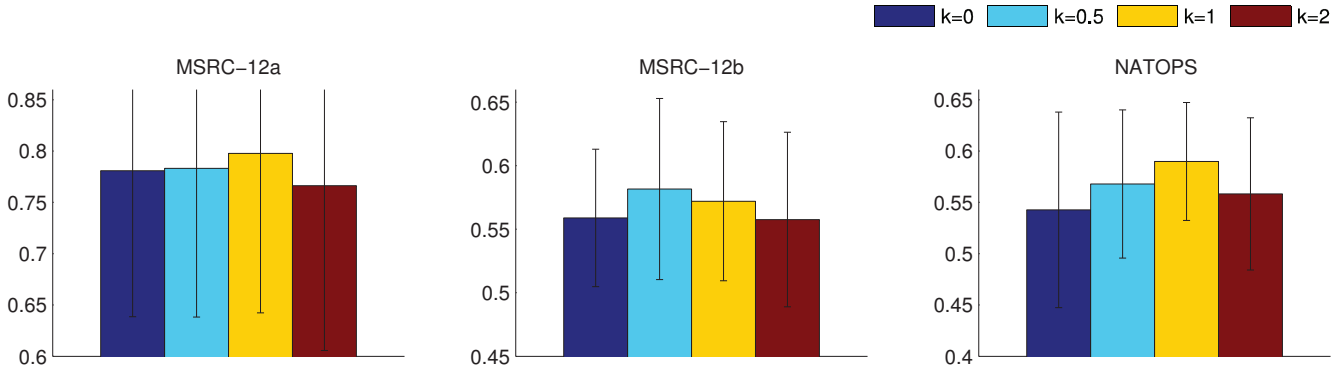

Fig. 4. Mean F1 scores as a function of k obtained from the Experiment II. See the text for details.

prior that deals with the imbalanced data problem. This puts more weight on the samples from underrepresented classes, allowing all samples in the dataset to have a balanced impact in the learning process. Experimental results on two recent gesture datasets, MSRC-12 and NATOPS, showed that, under the same amount of training data, the balanced data achieved better performance than the imbalanced data, indicating that learning from balanced data is important. We also showed that the distribution-sensitive prior improves the performance on the imbalanced data as well as the balanced data obtained using random undersampling.

Finding the optimal degree of the distribution-sensitive prior k (see Equation 3) is empirical and needs cross-validation at the current stage of our method. However, this can be automatically learned using meta-learning methods such as boosting approach. We plan to improve our approach by combining it with other techniques in imbalanced data learning literature.

VI. ACKNOWLEDGMENTS

This work was funded by ONR grant #N000140910625, NSF grant #IIS-1118018, NSF grant #IIS-1018055, and U.S.Army RDECOM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, USA, 1 edition, Jan. 1996.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 2002.

[4] A. Y. chung Liu. The effect of oversampling and undersampling on classifying imbalanced text datasets. Master's thesis, University of Texas at Austin, 2004.

[5] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 973–978, 2001.

[6] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*, pages 1737–1746, 2012.

[7] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, 21(9):1263–1284, 2009.

[8] T. Kanade, Y. li Tian, and J. F. Cohn. Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000)*, pages 46–53, 2000.

[9] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289, 2001.

[10] J. Laurikkala. Improving identification of difficult small classes by balancing class distribution. In *Artificial Intelligence Medicine, 8th Conference on AI in Medicine in Europe (AIME 2001)*, pages 63–66, 2001.

[11] C. X. Ling, Q. Yang, J. Wang, and S. Zhang. Decision trees with minimal costs. In *ICML*, 2004.

[12] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*, pages 57–64, 2011.

[13] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 1975–1981, 2010.

[14] D. Mease, A. J. Wyner, and A. Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8:409–439, 2007.

[15] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning (ICML 2004)*, 2004.

[16] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.

[17] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1848–1852, 2007.

[18] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[19] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[20] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[21] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*, pages 500–506, 2011.

[22] D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

[23] P. D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *J. Artif. Intell. Res. (JAIR)*, 2:369–409, 1995.

[24] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[25] G. Weiss and F. Provost. The effect of class distribution on classifier learning: An empirical study. Technical report, ML-TR-44, Department of Computer Science, Rutgers University, 2001.

[26] Y. Zhang and Z.-H. Zhou. Cost-sensitive face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(10):1758–1769, 2010.