RANSAC-based Training Data Selection for Speaker State Recognition

Elif Bozkurt[1]*, Engin Erzin*[1]*, Çiğdem Eroğlu Erdem*[2]*, A.Tanju Erdem*[3]

[1]Multimedia, Vision and Graphics Laboratory, Koç University, Istanbul, Turkey
[2]Department of Electrical and Electronics Engineering, Bahçeşehir University, Istanbul, Turkey
[3]Department of Electrical and Computer Engineering, Özyeğin University, Istanbul, Turkey
ebozkurt, eerzin@ku.edu.tr, cigdem.eroglu@bahcesehir.edu.tr, tanju.erdem@ozyegin.edu.tr

Abstract

We present a Random Sampling Consensus (RANSAC) based training approach for the problem of speaker state recognition from spontaneous speech. Our system is trained and tested with the INTERSPEECH 2011 Speaker State Challenge corpora that includes the Intoxication and the Sleepiness Sub-challenges, where each sub-challenge defines a two-class classification task. We aim to perform a RANSAC-based training data selection coupled with the Support Vector Machine (SVM) based classification to prune possible outliers, which exist in the training data. Our experimental evaluations indicate that utilization of RANSAC-based training data selection provides 66.32 % and 65.38 % unweighted average (UA) recall rate on the development and test sets for the Sleepiness Sub-challenge, respectively and a slight improvement on the Intoxication Sub-challenge performance.

Index Terms: Speaker State Challenge, Intoxication, Sleepiness, RANSAC

1. Introduction

For supervised pattern recognition problems large training sets need to be recorded and labeled to be used for the training of the classifier. The labeling of large datasets is a tedious job, carried out by humans and hence prone to human mistakes. The mislabeled (or noisy) examples of the training data may result in a decrease in the classifier performance. It is not easy to identify these contaminations or imperfections of the training data since they may also be "hard to learn examples". In that respect, pointing out troublesome examples is a "chicken-and-egg" problem, since good classifiers are needed to tell which examples are noisy [1]. In this work, we assume that outliers in the training set of speech recordings mainly result from mislabeled or ambiguous data. Our goal is to remove such noisy samples from the training set to increase the performance of support vector machine (SVM) classifiers on the INTERSPEECH 2011 Speaker State Challenge corpora.

2. Related Work

Previous research on data cleaning, which is also called as data pruning or decontamination of training data shows that removing noisy samples is worthwhile [1] [2] [3]. Guyon et al. [4] have studied data cleaning in the context of discovering informative patterns in large databases. They mention that informative patterns are often intermixed with unwanted outliers, which are errors introduced non-intentionally to the database. Informative patterns correspond to atypical or ambiguous data and are pointed out as the most "surprising" ones. On the other hand, garbage patterns are also surprising, which correspond to meaningless or mislabeled patterns. The authors point out that automatically cleaning the data by eliminating patterns with suspiciously large information gain may result in loss of valuable informative patterns. Therefore they propose a user-interactive method for cleaning a database of hand-written images, where a human operator checks those patterns that have the largest information gain and therefore the most suspicious.

Batandela and Gasca [2] report a cleaning process to remove suspicious instances of the training set or correcting the class labels and keep them in the training set. Their method is based on the Nearest Neighbor classifier. Wang et al. [5], present a method to sample a large and noisy multimedia data. Their method is based on a simple distance measure that compares the histograms of the sample set and the whole set in order to assess the representativeness of the sample set. The proposed method deals with noise in an elegant way, and has been shown to be superior to the simple random sample (SRS) method [6][7].

Angelova et al. [1] present a fully automatic algorithm for data pruning, and demonstrate its success for the problem of face recognition. They show that data pruning can improve the generalization performance of classifiers. Their algorithm has two components: the first component consists of multiple semi-independent classifiers learned on the input data, where each classifier concentrates on different aspects and the second component is a probabilistic reasoning machine for identifying examples which are in contradiction with most learners and therefore noisy.

There are also other approaches for learning with noisy data based on regularization [8] or averaging decisions of several functions such as bagging [9]. However, these methods are not successful in high-noise cases.

3. Contribution and Outline of the Paper

In this paper, we adopt RANSAC-based data selection for training SVM classifiers for the speaker state recognition problem. RANSAC is a paradigm for fitting a model to noisy data and utilized in many computer vision problems [10]. RANSAC performs multiple trials of selecting small subsets of the data to estimate the model. The final solution is the model with maximal support from the training data. The method is robust to considerable amount of outliers. Aside from data pruning, RANSAC has also been used for classifier parameter selection on large datasets [11].

The outline of the paper is as follows. In Section 4, background information is provided for the well known RANSAC algorithm. In Section 5, the proposed RANSAC-based data

selection approach is described. In Section 6, our experimental results are provided, which is followed by conclusions and future work given in Section 7.

4. The RANSAC algorithm

Random Sample Consensus (RANSAC) is a method for fitting a model to noisy data [12]. RANSAC is capable of being robust to error levels of significant percentages. The main idea is to identify the outliers as data samples with greatest residuals with respect to the fitted model. These can be excluded and the model is re-computed over the consensus set. The consensus set, also called inlier set refers to final outlier cleaned sample set. The steps of the general RANSAC algorithm are as follows [10] [12]:

1. Suppose we have n training data samples $X = x_1, x_2, ..., x_n$ to which we hope to fit a model determined by (at least) m samples ($m \leq n$).
2. Set an iteration counter $k = 1$.
3. Choose at random m items from X and compute a model.
4. For some tolerance ε, determine how many elements of X are within ε of the derived model. If this number exceeds a threshold t, re-compute the model over this consensus set and stop.
5. Set $k \leftarrow k + 1$ If , $k < K$ for some predetermined K, go to 3. Otherwise, accept the model with the biggest consensus set so far, or fail.

There are possible improvements to this algorithm [10] [12]. The random subset selection may be improved if we have prior knowledge of data and its properties, that is some samples may be more likely to fit a correct model than others.

There are three parameters that need to be chosen:

- The acceptable deviation from a good model: ε. It is empirically determined by fitting a model to m points, measuring the deviations and setting to some number of standard deviations above the mean error.
- The size of the consensus set: t. It should represent sample points for a sufficient model and number of samples to refine the model to the final best estimate. For the sufficient model a value of t satisfying $t - m > 5$ has been suggested [12].
- The maximum iteration count: K. Values of $K = 2\omega^{-m}$ or $K = 3\omega^{-m}$ have been argued to be reasonable choices [12], where ω is the probability of a randomly selected sample to be within ε of the model.

5. RANSAC-based Data Selection

5.1. Speech Features and the Classifier

We use the official IS 2011 SSC feature set provided by the challenge organizers per corpus. The feature set consists of 4368 features those are built from three sets of low-level descriptors (LLD) and one corresponding set of functionals for each LLD set [13]. As for the classifier, we use Support Vector Machines (SVMs) with linear kernel for learning. The SVM implementation we use is the LIBSVM toolkit [14].

5.2. RANSAC-based Training of SVM Classifiers

Our classifier training strategy is composed of two stages: in the first stage, we draw an initial approximation of outliers by using binary SVM models in combination with the RANSAC algorithm. Based on the initial approximation we apply a second stage classification on the inlier data with supervised training using binary SVMs that maximizes the margin.

Our goal is to train SVM classifiers for each of the subchallenges (Sleepy vs. non-sleepy and alcoholized vs. non-alcoholized). For each class, we want to select a training set such that the fraction of the number of inliers (consensus set) over the total number of utterances in the training dataset is maximized. For determining the biggest consensus set (inliers) for each of the classes, we use the linear kernel SVM structure with appropriate hyperparameters.

Initially, we randomly select a training data subset (of size m) for each class in the first classification stage. Then, we train our model over the initial random set (of size $2m$) and test on the remaining samples from the training set. The decion whether a single element is an outlier or not is based on the SVM recognition result. The well-known RANSAC algorithm considers the number of inliers above a threshold t as a metric to determine the model with highest support from the training data, as described in Section 4. However, for our case if we were to evaluate subsets considering the highest number of inliers, we would have favored the WA recall rate. Since the primary evaluation criterion of the challenge is UA recall rate, we seek a random subset that would achieve the highest UA recall rate on the remaining training set instances. After the RANSAC-based training data cleaning method, we aim to obtain a training set with higher discrimination ability than the provided one. Therefore, we consider both classes during the outlier cleaning process and instead of using number inliers, we use the UA recall rate as the threshold value t in our experiments.

The steps of the RANSAC-based SVM training method are as follows:

1. Suppose a class in the training set has n training data samples $X = x_1, x_2, ..., x_n$ to which we hope to fit a model determined by (at least) m samples ($m \leq n$). Randomly select m samples from each class.
2. Set an iteration counter k = 1.
3. Train the binary SVM classifier with a linear kernel based on the randomly selected subset of size $2m$.
4. For each class determine how many elements of remaining samples of training set are identified as inliers based on the tolerance value ε. Then, calculate the UA recall rate: if this rate exceeds a threshold t, recompute the model over this consensus set.
   Initially, set $t = 0$. Then, use UA recall rate of the previously selected model as the threshold value t. The tolerance value ε corresponds to small error penalty in the SVM definition in our case.
5. Increase the iteration counter $k \leftarrow k + 1$, If $k < K$, and $k < 2000$, for some predetermined K, go to step 3. Otherwise, accept the model with the biggest consensus set so far, or fail. Here, we estimate K, the number of loops required for the RANSAC algorithm to converge, using the number of inliers [9]:

$$K = \frac{\ln(1-p)}{\ln(1-\omega^m)} \tag{1}$$

Here we set $\omega = \frac{m_i}{m}$, where m_i is the number of inliers for iteration i and $p = 0.99$ is the probability that at least one of the sets of random samples does not include an outlier.

The first stage of the classification process, namely RANSAC-based training data cleaning method with SVMs will try to achieve a more coherent inlier set at every iteration since, we update the threshold value t in case, we achieve a better UA recall rate on the remaining samples of the training set. Then, in the second stage, the model with the highest support from the training set is used for the classification of the speaker states. We again use a basic SVM with a linear kernel that is a binary classifier which seeks to fit an optimal separating hyperplane or decision boundary between the classes.

6. Experimental Results

In this section, we present our experimental results for the two-class speaker state recognition problem using the Intoxication and the Sleepiness Sub-challenge databases provided by the INTERSPEECH 2011 Speaker State Challenge organizers [13]. The distribution of speaker state classes in the databases is highly unbalanced so that the primary performance evaluation measure is the unweighted average (UA) recall rate which is the average recall rate of the two classes.

We use SVMs as the classifier in our experiments. Initially, we scale the dataset since scaling before applying SVM is very important. The main advantage of scaling is to avoid feature values in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Then, we apply our RANSAC-based data cleaning approach coupled with SVMs.

Our RANSAC-based data cleaning approach randomly selects subsets of size m as mentioned in sub-section 5.2. We vary the parameter m in the range from 200 to 600 by 200 for both of the Intoxication and the Sleepiness Sub-challenges. UA and WA recall rates on the development set vs. RANSAC-based training subset size (m) relationship is shown in Figure 1. For both of the sub-challenges the UA and WA recall rates on the development set are highest when $m = 200$ subset size is selected. The number of samples determined as inliers after the RANSAC-based training data cleaning process is given in Table 1 for the sub-challenges when $m = 200$. Samples detected as outliers are excluded from the training data for the sake of obtaining higher recognition rates on the development and test sets. For the Intoxication and Sleepines Sub-challenges highest UA recall rates on the development sets are 62.25 % and 66.32 % when $m = 200$. As we increase m further, the performance of the system on the development set decreases. The reason for this situation may be overfitting of the linear kernel SVM models.

We list the UA and WA (weighted average) recall rates for the RANSAC-based training of linear kernel SVM classifiers in comparison to recognition results using all the available training data, in Table 2. Results in the table are given for training on the train partition and testing on the development partition for each sub-challenge. The UA and WA classification rates for the two-class classification task of the Intoxication Sub-challenge are 62.43 % and 68.98 %, respectively when all the available training data is modeled with linear kernel SVM classifiers. Using RANSAC-based data cleaning method with random subset size $m = 400$, increases the performance up to 62.50 % UA and 64.74 % WA recall rates, respectively.

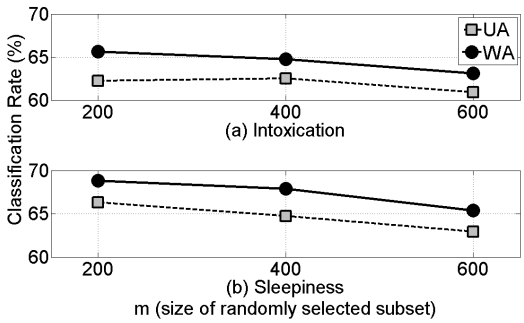

Figure 1: *Unweighted and weighted average recall rate on the development set vs. RANSAC-based training random subset size relationship for the Intoxication and the Sleepiness Sub-challenges.*

Table 1: *Intoxication and Sleepiness Sub-Challenge number of samples for the whole and RANSAC-cleaned training sets when* $m = 200$.

	Intoxication Sub-challenge	
Training set	Whole set	RANSAC-cleaned
NAL	3750	2701
AL	1650	1301
Total	5400	4002
	Sleepiness Sub-challenge	
Training set	Whole set	RANSAC-cleaned
NSL	2125	1620
SL	1241	996
Total	3366	2616

Similarly, the UA and WA classification rates on the development set for the Sleepiness Sub-challenge are 61.78 % and 65.45 %, respectively when all the available training data is modeled with linear kernel SVM classifiers. Using RANSAC-based data cleaning method with random subset size $m = 200$, increases the performance up to 66.32 % UA and 68.81 % WA recall rates, respectively.

Moreover, we achieve better recognition performance results on the Sleepiness Sub-challenge test set when linear kernel SVM models are trained on RANSAC-cleaned training data. As shown in Table 3, on the test set we achieve 65.38 % UA and 71.97 % WA recall rates with our proposed method. If we use all the available training data (excluding the development set) for training models, we obtain 63.86 % UA and 70.55 % WA recall rates on the test set. Table 4 is the confusion matrix for our RANSAC-based data cleaning approach result on the test set of the Sleepiness Sub-challenge.

As seen in Table 2, RANSAC-based training attains higher classification rates than the full SVM with much smaller model sizes. Number of support vectors for the Intoxication and Sleepiness Sub-challenges are 2489 and 1329, respectively when whole training set is used for training the linear kernel SVMs. In case RANSAC-based data cleaning with random subset size $m = 200$ is used, the number of support vectors decreases to 844 and 581, respectively. Thus, our approach can improve generalization performance and lower storage requirements, as well.

Nevertheless, our recognition rates for both of the sub-

Table 2: *Intoxication and Sleepiness Sub-Challenge results on the development set by unweighted and weighted accuracy (UA/WA). Binary SVM with linear kernel is applied to whole and RANSAC-cleaned training set. The official feature set, IS 2011 SSC of the challenge is used for different random subset sizes, m.*

Sub-challenge	**Intoxication**		**Sleepiness**	
	%UA	%WA	%UA	%WA
IS 2011 SSC baseline	65.30	69.20	67.30	69.10
Whole training set	62.43	68.98	61.78	65.45
RANSAC ($m = 200$)	62.25	65.63	**66.32**	**68.81**
RANSAC ($m = 400$)	62.50	64.74	64.74	67.89
RANSAC ($m = 600$)	60.93	63.10	62.93	65.38

Table 3: *The Sleepiness Sub-Challenge results on the test set by unweighted and weighted accuracy (UA/WA). Binary SVM with linear kernel is applied to whole and RANSAC-cleaned training set. The official feature set, IS 2011 SSC of the challenge is used for random subset size $m = 200$.*

	Sleepiness	
Training set	%UA	%WA
IS 2011 SSC baseline	70.30	73.00
Whole training set	63.86	70.55
RANSAC ($m = 200$)	65.38	71.97

challenges are below the baseline results although we also use linear kernel SVMs classifiers. The main reason for this outcome is that we do not use the Synthetic Minority Over-sampling Technique (SMOTE) in our experiments as the challenge organizers do [13]. SMOTE method generates synthetic instances on the basis of nearest neighbour approach to handle the class imbalance problem.

7. Conclusions and Future Work

In this paper, we presented a random sampling consensus based training data selection method for the problem of speaker state recognition. The experimental results show that the proposed method is promising for SVM based speaker state recognition from spontaneous speech data in the Sleepiness sub-challenge. We get 66.32 % UA and 68.81 % WA recall rates on the development set. On the test set we achieve 65.38 % UA and 71.97 % WA recall rates. For the the Intoxication Sub-challenge the proposed approach does not degrade performance when $m = 400$. However, we do not gain much performance results either. We conclude that few outliers exist in the Intoxication Sub-challenge training set but, the Sleepiness Sub-challenge training set has outliers. RANSAC-based training data selection approach eliminates outliers in the Sleepiness Sub-challenge training set and improves performance compared to using all the available training data (excluding the development set).

As the distribution among classes is not balanced, Synthetic Minority Over-sampling Technique (SMOTE) can be used to balance the instances in the respective learning partitions after RANSAC-based training data selection approach as a future work. In order to increase the benefits of the data cleaning approach, and to decrease the training effort, the algorithm may be improved by using semi-deterministic subset selection methods.

Table 4: *Confusion matrix for the Sleepiness Sub-Challenge results on the test set. Binary SVM with linear kernel is trained with RANSAC-cleaned training set. The official feature set, IS 2011 SSC of the challenge is used for random subset size $m = 200$.*

	NSL	SL	Sum
NSL	1607	350	1957
SL	437	414	851

8. Acknowledgements

We would like to thank the INTERSPEECH 2011 Speaker State Challenge team for their initiative and for kindly providing the challenge database and test results. This work was supported in part by the Turk Telekom and the Turkish Scientific and Technical Research Council (TUBITAK) under projects 106E201, 110E056 and COST2102 action.

9. References

[1] A. Angelova, Y. Abu-Mostafa, and P. Perona, "Pruning training sets for learning of object categories," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[2] R. Barandela and E. Gasca, "Decontamination of training samples for supervised pattern recognition methods," *Lecture Notes in Computer Science*, vol. 1876, pp. 621–630, 2000.

[3] I. Ben-Gal, *Outlier Detection, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005.

[4] I. Guyon, N. Matin, , and V. Vapnik, "Discovering informative patterns and data cleaning," in *Workshop on Knowledge Discovery in Databases*, 1994.

[5] S. Wang, M. Dash, L. Chia, and M. Xu, "Efficient sampling of training set in large and noisy multimedia data," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, 2007.

[6] B. Gu, F. Hu, and H. Liu, "Sampling and its applications in data mining: A survey," *Tech. Rep. School of Computing, National University of Singapore*, 2000.

[7] F. Olken, *Random Sampling from Databases*. Ph. D. Thesis, Department of Computer Science, University of California, Berkeley, 1993.

[8] G. Ratsch, T. Onada, and K. Muller, "Regularizing adaboost," *Advances in Neural Information Processing Systems*, vol. 11, pp. 564–570, 2000.

[9] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[10] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*. Thomson, 2008.

[11] K. Nishida and T. Kurita, "Ransac-svm for large scale datasets," in *International Conference on Pattern Recognition*, Florida, USA, 2008.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Graphics and Image Processing*, vol. 24, 1981.

[13] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Kraajewski, "The Interspeech 2011 speaker state challenge," in *Interspeech (2011), ISCA*, Florence, Italy, 2011.

[14] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.