An Agreement and Sparseness-based Learning Instance Selection and its Application to Subjective Speech Phenomena

Zixing Zhang[1], **Florian Eyben**[1], **Jun Deng**[1], **and Björn Schuller**[2,1]

[1] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2] Department of Computing, Imperial College London, United Kingdom
zixing.zhang@tum.de

Abstract

Redundant instances in subjective speech phenomena may cause increased training time and performance degradation of a classifier like in other pattern recognition tasks. Instance selection, aiming at discarding some 'troublesome' instances and choosing the most informative ones, is a way to solve this issue. We thus propose a tailored algorithm based on human *A*greement levels of labelling and class *S*parseness for learning *I*nstance *S*election – ASIS for short. Extensive experiments on a standard speech emotion recognition task show the effectiveness of ASIS, indicating that by selecting only 30 % of the training set, the system performance significantly outperforms training on the whole training set without instance balancing. In terms of performance it remains comparable to the classifier trained with instance balancing, but at a fraction of the training material.

Keywords: Instance Selection, Subjective Speech Phenomena, Human Agreement Level, Sparse Instance Tracking

1. Introduction

Instance selection is important in many pattern recognition tasks. This includes in particular also the field of automatic recognition of (often highly) *subjective* paralinguistic speech phenomena, such as speakers' emotion, interest, sleepiness, intoxication, or voice likability. There are three main reasons why instance selection is worth considering: Following the idea of "there is no data like more data", many efforts have recently been undertaken to collect and/or create large amounts of data with the aim to improve recognition performance: more manual annotations, aggregation of multiple corpora (Schuller et al., 2011), and semi-supervised learning or co-training (Zhang et al., 2011; Zhang et al., 2013). However, as the size of the data set which is used to train a classifier increases, the complexity of the models and the training time increases (Schuller et al., 2012b). Even though, for most commercial applications classifiers training is done once and is not a time critical operation, faster training times gives companies a competitive advantage. Researchers, on the other hand, will train many models when optimising parameters and testing new methods. Thus, they largely benefit from reduced training times.

Another main reason is the subjectivity of the paralinguistic phenomena. Unlike traditional pattern recognition tasks where a true 'ground truth' is available, those tasks only have 'gold standard' labels, which are often assigned by (sometimes weighted) majority voting over multiple human ratings. In fact, instance labelling for such tasks highly depends on the labellers' personal judgement. For music mood, for example, some would consider a musical piece more sad or happy than others or even have opposing views due to personal associations with a song. The same holds for speaker emotion or likability recognition (cf. (Sneddon et al., 2012; Schuller, 2013)). Instances with high labeller uncertainty could potentially cause the model to over-fit these 'noisy' instances resulting in increased complexity (Angelova, 2004). This thus would deteriorate the generalization performance.

The last reason relates to the imbalance of the number of instances among classes, which is most pronounced in databases with natural and spontaneous speech, where 'neutral' speech is much more frequent than clear cut cases of emotional or other target speech. This leads to the fact, that some models tend to favour the majority classes and thus show a bad performance on the sparse (minority) classes. However, these sparse classes are usually of most interest in practical applications.

Therefore, a reduction of the amount of training instances is beneficial if the following two criteria are met: 1) Equal or improved performance. That is, the model trained on a subset should perform equally to or better than the model trained on all instances; and 2) Reduction of training time. To this end, we propose an instance selection method in this paper which is based on human labeller agreement level and class sparseness. The two main contributions of this paper are: 1) We investigate whether pruning of the instances with the lowest labeller agreement improves performance; and 2) After pruning we select an equal amount of instances from each class in order to produce a set with a balanced number of instances of each class.

1.1. Related Work

In pattern recognition, numerous methods have been proposed and investigated in the literature for solving the data selection problem. Most of them can be assigned to one of the following two groups from a technical point of view (Liu and Motoda, 2002; Olvera-López et al., 2010):

The first group is *wrapper-based selection*, where the selection criterion is based on the accuracy obtained by a classifier (Olvera-López et al., 2010). Those instances that do not improve predictive performance of classification will be discarded from the training set. Most of the wrapper-based selection methods are related to the k-nearest neighbour classifier (Cover and Hart, 1967) like the Condensed

Nearest Neighbour (CNN) (Hart, 1968), Selective Nearest Neighbour rule (SNN) (Ritter et al., 1975), or Decremental Reduction Optimisation Procedure (DROP) (Wilson and Martinez, 2000). With CNN, for example, the instances misclassified by the classifier will be selected and added into the initial training set.

Unlike wrapper-based selection methods, *filter-based selection* methods in the second group attempt to select the instances by means of sampling or clustering, without depending on a prediction of a classifier (Olvera-López et al., 2010). Among them, a prominent algorithm is RANdom SAmple Consensus (RANSAC) proposed by Fishler and Bolles (Fischler and Bolles, 1981). It uses a data set as small as possible to determine model parameters – mostly used for estimating homography transformation matrices in computer vision. Then, other data are tested against the estimated model and those data which fit the model within a predefined tolerance ϵ will be considered part of a consensus set. Whenever the ratio of the number of consensus data to the total number data in the set exceeds a predefined threshold, the model parameters are re-estimated using all consensus and all initial data. This procedure is repeated a fixed number of times. Another example is the Pattern by Ordered Projections (POP) (Riquelme et al., 2003) which discards interior instances and selects some border instances, where a border instance is defined by its nearest neighbour belonging to other classes, and an interior instance is defined by its nearest neighbour belonging to the same class. In addition, to address the issue of class imbalance, e. g., Garcia et al. proposed a scalable instance selection method in (García-Pedrajas et al., 2013).

However, most of these methods are developed for objective pattern recognition tasks with a definite ground truth, such as face recognition (Angelova et al., 2005), textual news classification into groups (Fragoudis et al., 2002), speech recognition, or language translation (Wu et al., 2007; Lu et al., 2012). Even though there is some work dealing with subjective pattern recognition tasks (e. g., (Erdem et al., 2010) which selects a training subset by RANSAC for emotion recognition), the influence of labelling uncertainty on recognition performance has not been considered directly nor has the class imbalance problem been addressed. These two issues are the focus of the work presented in this paper.

In the following, we introduce the details of our proposed instance selection algorithm in Section 2.. Then, we describe the databases used for the experiments and discuss the results of the proposed instance selection algorithm in Section 3.. Finally, we draw the conclusions in Section 4..

2. Methodology

The main idea of our algorithm is to discard the instances with low labelling agreement and afterwards sub-sample the data set by selecting an equal amount of instances for each class from the remaining instances.

2.1. Human Agreement Levels

To measure human inter-rater agreement levels, we employ Fleiss' frequently used Kappa coefficient, which is expressed as:

$$\kappa := \frac{p_0 - p_c}{1 - p_c}, \tag{1}$$

where p_0 is the observed agreement of labellers, and p_c is chance-level agreement. In the case of a single instance, the probability of p_0 can be simplified by estimating the proportion of cases in which labellers agree on a common category:

$$p_0 = \sum_{m=1}^{M} \frac{\eta_m}{M}, \tag{2}$$

where $\eta_m \in (0, 1)$ stands for a binary annotation of a specific category, and M is the number of labellers. Thus, the difference $p_0 - p_c$ indicates the proportion of cases where 'beyond-chance agreement' occurs. It is normalized by the probability of disagreement $1 - p_c$ which is expected by chance.

2.2. ASIS: Agreement and Sparseness-based Instance Selection

The details of the proposed algorithm are presented in Algorithm 1. It includes two steps: Agreement-based Instance Selection (AIS) and Sparseness-based Instance Selection (SIS).

The AIS step aims at discarding the most noisy instances mainly caused by high disagreement of human labelling. In this process, we prefer to *proportionally* discard instances across classes. On the one hand, it prevents the case of potential maldistribution of instances which might result in discarding such instances mainly belonging to certain classes, especially the sparse ones. On the other hand, it probably improves the separability of classes by potentially removing instances close to the class boundary in the feature space. Therefore, the larger we choose the discarded subset ($P_D[\%]$) to be, the fewer instances – relatively seen – might be located near the class boundaries, and the less complex the model becomes.

The SIS step randomly selects an equal number of instances from each class set with the aim of coping with the class sparseness problem. Note, that in the case of a class balanced task the size of the selected subset ($P_S[\%]$) will satisfy $P_D + P_S \leq 1$, while in the case of a class imbalanced task, P_S is limited by the instance count of the most sparse class. This problem could be eased to some extent by loosening the constraint of ending up with a balanced distribution of instances after sub-sampling, i. e., the missing amount of instances of the sparse classes can be filled in with instances from the abundant classes. In this paper, however, we adhere to the strict rule of balanced selection and evaluate binary tasks only as straightforward examples.

3. Experiments

To evaluate the effectiveness of our algorithm, we selected two well-standardised machine learning tasks and according data from the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009) and the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012a). Both are of highly subjective nature and together they cover the spectrum from short-term (emotion) to long-term (likability) speaker traits. In the following, the two according

Algorithm 1: ASIS: The proposed agreement and sparseness-based instance selection algorithm.

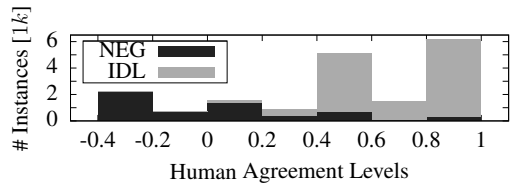
Input:
$\mathcal{D}$: Database of N instances annotated in classes C_i $(i = 1, \cdots, k)$ and corresponding human agreement levels l;
P_D: Size of discarded subset with low human agreement levels (percentage of the full training set);
P_S: Size of selected subset (percentage of the full training set);
k: number of classes;
Output:
$\mathcal{S}$: Subset of database $\mathcal{D}$;

Process
Obtain the proportional distribution of each class R_i $(i = 1, \cdots, k)$ in the training set of $\mathcal{D}$;
(Step: Agreement-based Instance Selection (AIS))
for $i = 1, \cdots, k$ **do**
    Sort the instances that are annotated as class C_i by human agreement levels l from low to high, producing queue Q_i;
    Delete $n_{Di} = N \times P_D \times R_i$ instances which are at the beginning of Q_i;
end
(Step: Sparsness-based Instance Selection (SIS))
for $i = 1, \cdots, k$ **do**
    Randomly select $n_{Si} = N \times P_S / k$ instances belonging to class C_i;
end
Fuse n_{Si} $(i = 1, \cdots, k)$ into one output subset $\mathcal{S}$.

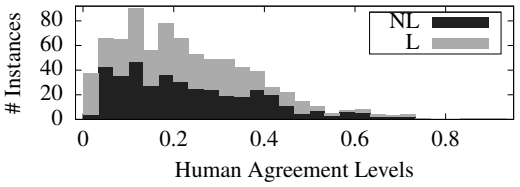
databases are introduced and then the results obtained on these sets are described.

3.1. Emotion: FAU Aibo Emotion Corpus

The FAU Aibo Emotion Corpus (Steidl, 2009) is the official corpus of the INTERSPEECH 2009 Emotion Challenge (EC) (Schuller et al., 2009). It contains recordings of children interacting with Sony's pet robot Aibo. The language is German. The Wizard-of-Oz controlled Aibo robot sometimes disobeyed children's commands, thereby provoking various emotional reactions. The recording was done at two different schools – MONT and OHM –, and features 51 children with 21 boys and 30 girls at ages ranging from 10 to 13 years. Five labellers listened to the turns in sequential order and labelled each word independently from each other as neutral or as belonging to one of ten other emotion classes. In the challenge, the final labelling and human agreement levels for chunks are determined by majority voting on labels of the five labellers on the word level onto one label for the whole chunk. Then, chunks were grouped into the 2-class labelling: **NEG**ative (subsuming $angry$, $touchy$, $reprimanding$, and $emphatic$) and **IDL**e (consisting of all other states). Fig. 1 (a) shows the instance distribution of the training set with human agreement levels. For our experiments, we use the whole corpus consisting of 18 216 chunks, where the training set includes 3 358 'NEG' and 6 601 'IDL' instances, and the test set consists of 2 465 'NEG' and 5 792 'IDL' instances. Note that, for the sake of balancing categories, some instances with negative human agreement level also belong to the class 'NEG'.



(a) Emotion: AEC



(b) Likability: SLD

Figure 1: Number of instances with human agreement levels in the AEC (a), and SLD (b).

3.2. Likability: Speaker Likability Database

The Speaker Likability Database (SLD) (Burkhardt et al., 2010) was the official corpus of the Likability Sub-Challenge in the INTERSPEECH 2012 Speaker Trait Challenge (STC) (Schuller et al., 2012a). The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. An age and gender balanced set of 800 speakers is selected. For each speaker, the longest sentence (consisting of a command embedded in a free sentence) was selected. Likability rating was executed by 32 labellers according to how well they personally liked the voices. They were asked not to take into account the linguistic content or the transmission quality. The rating was done on a seven point Likert scale ([-3, -2, -1, 0, 1, 2, 3]). To establish a coherent consensus from the highly individual likability ratings, the evaluator weighted estimator (EWE) (Grimm and Kroschel, 2005) was used in the challenge. It uses higher weights for more agreeable labellers. Based on the median EWE rating of all stimuli in the SLD, the data was discretised at the threshold of 0.108 into the classes—'likable' (**L**, $EWE > 0.108$) and 'non-likable' (**NL**, $EWE < 0.108$). The final challenge set of 800 instances is partitioned as follows: training set (L, 189; NL, 205), development set (L, 92; NL, 86), and test set (L, 119; NL, 109). Fig. 1 (b) shows the instance distribution along with human agreement levels. Here, we slightly modify Kappa as follows:

$$Adapted \quad Fleiss' \quad \kappa := \left| \frac{p_{ewe} - p_t}{p_{ewe_{max}} - p_t} \right|, \qquad (3)$$

by replacing the p_c with the threshold of 'L' and 'NL' p_t at 0.108, p_o with the EWE values p_{ewe}, and 1 with the maximum EWE value of p_{ewe}. Therefore, the instances with an EWE value near 0.108 are considered as low agreement and vice versa.

3.3. Protocol and Results

As in the challenge tasks, we evaluate performance in terms of unweighted average recall (UAR). In addition, we use the original challenge feature sets for the tasks of emotion and likability recognition in our experiments. Thus, for emotion recognition, we use 384 features resulting from a systematic combination of 16 low-level-descriptors (LLDs) and corresponding first order delta coefficients with 12 functionals (Schuller et al., 2009); for likability recognition, we utilize 6 125 features by brute-forcing based on 64 LLDs and 61 functionals (Schuller et al., 2012a) – all features are extracted with the open-source toolkit openSMILE (Eyben et al., 2010). In the same vein, we keep the classifiers, their implementations, and parameters as in Challenges: for emotion recognition, Support Vector Machines (SVMs) trained by Sequential Minimal Optimization (SMO) with polynomial kernel (degree 1) and a complexity constant of 0.05; for likability recognition, Random Forests (RF) with a number of trees $N = 1\,000$ and a feature subspace size of $P = .02$. The Weka toolkit (Hall et al., 2009) is used in both cases. Note, that the instance selection algorithm is only applied on the training set. The test set is not modified and kept the same as in the original Challenge setup in order to allow for a direct comparison.

3.4. Emotion

The following experiments were executed for emotion recognition with different variations of the instance selection algorithm: 1) only agreement-based instance selection ('AIS') based on discarding low-agreement instances (cf. Step 'AIS' in Algorithm 1); 2) only sparseness-based instance selection ('SIS') by selecting sparse instances (cf. Step 'SIS' in Algorithm 1); 3) both steps (ASIS) at the same time (random selection with balancing of instances across classes).

For comparison, we denote the control methods of Random Instance Selection (RIS) as randomly selecting a predefined number of instances from the whole set without other constraints.

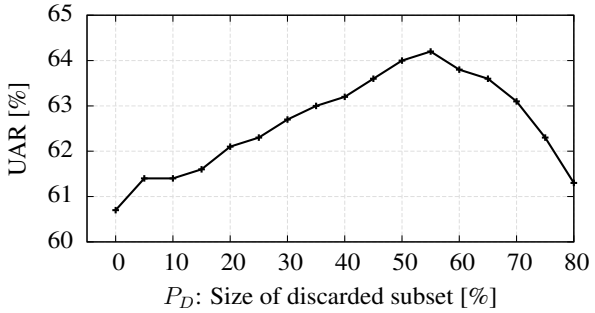


Figure 2: Agreement-based Instance Selection (AIS): UAR on the AEC test set after discarding low agreement training instances (no balancing).

Fig. 2 gives an overview on performances after discarding a certain ratio of instances with low human agreement (AIS). Note, that the human agreement levels by discarding 5, 10, 20, 30, 40, 50, 60 % of the instances for the class IDL are

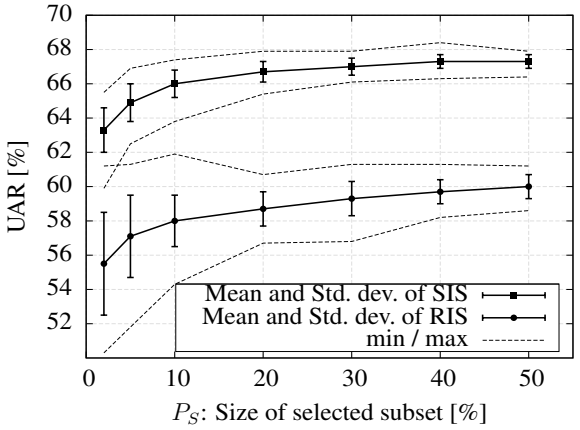


Figure 3: Sparseness-based Instance Selection (SIS): UAR mean, standard deviation (std. dev.), minimum (min), and maximum (max) on the AEC test set over 40 independent runs. Comparison of balanced SIS and random instance selection (RIS) from the training set. No discarding of instances with low agreement ($P_D = 0$).

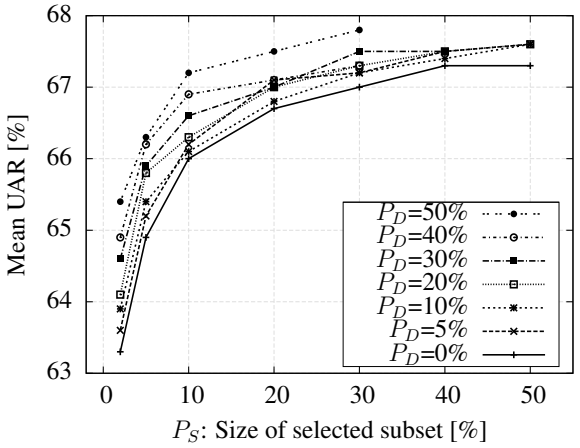


Figure 4: Agreement and Sparseness-based Instance Selection (ASIS): Mean UAR on the AEC test set in 40 independent runs of *balanced sub-sampling* after discarding P_D instances with lowest labeller agreement from the training set of the AEC.

0.4, 0.52, 0.60, 0.60, 0.60, 0.72, 0.90, and for the class NEG are -0.28, -0.2, -0.2, -0.2, -0.08, 0.06, 0.2, respectively. No instance balancing is performed here. The performance of the classifier improves continuously and significantly (one-sided z-test) until 55 % of the training set instances with human agreement are discarded (from 60.7 % to 64.2 % UAR). Fig. 3 compares the performance of two instance sub-sampling strategies (with (SIS) and without balancing), both without any prior discarding of low agreement instances. As expected, UAR is increased by about 8 % absolute when balancing is performed, showing the importance of a balanced distribution for SVM (and further) classifiers. Fig. 4 shows results obtained when randomly sub-sampling the training set and balancing after discard-

ing low agreement instances (ASIS). At a certain ratio of discarded instances, increasing the number of selected instances enhances the system robustness. As more instances are added, however, the increase of UAR converges. At a certain amount of sub-sampling, discarding up to 50 % of low agreement instances improves UAR. Note that this improvement is more obvious for a small subset size, as in this case the disturbing influence of the low agreement instances has a larger relative impact on the model. The best result of 67.8 % of UAR is achieved by discarding 50 % of lowest agreement instances and selecting only 30 % of instances (relative to the whole set) for model building. This is equivalent to the baseline (67.7 % of UAR) in (Schuller et al., 2009) where the whole training set with Synthetic Minority Oversampling TEchnique (SMOTE) is considered (for balancing). Note that, for this experiment the amount of sub-sampling is limited by the size of the minority class 'NEG'.

3.5. Likability

We further evaluate the potential of our algorithm for a secondary task: automatic speaker's voice likability recognition. Fig. 5 visualises the performance after discarding instances with low agreement levels (AIS). Table 1 shows the relationship between the percentage of discarded instances and the human agreement levels of the classes 'L' and 'NL'. Due to the way the classes 'L' and 'NL' have been defined (by median), the instances are already balanced among the two classes. Thus, no balancing is therefore necessary (i. e., no SIS). By discarding the lowest 10 % agreement levels, the UAR is raised from 59.0 % to 62.0 %. One notices that discarding more instances does not bring additional improvement. This might be due to the small size of the dataset with only 600 instances in the training set.

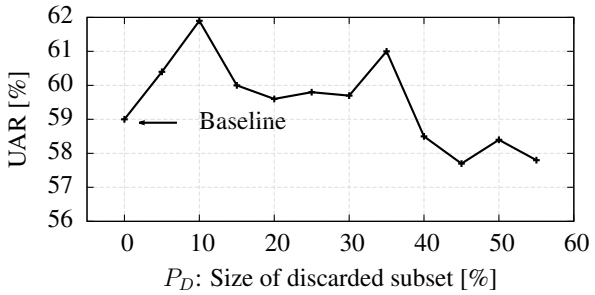


Figure 5: Agreement-based Instance Selection (AIS): UAR on the SLD (likability) test set after discarding low agreement training instances from the training set.

Table 1: Relationship percentage of discarded instances and agreement levels

	Percentage discarded				
Levels	10 %	20 %	30 %	40 %	10 %
L	0.01	0.05	0.09	0.15	0.18
NL	0.07	0.10	0.13	0.17	0.20

4. Conclusions

We proposed ASIS – agreement and sparseness-based instance selection which exploits labeller agreement levels and the concept of sparse class learning by random subsampling of the training space. We demonstrated the potential of this algorithm for two standard machine learning challenge tasks for speech emotion and voice likability recognition. For the emotion recognition experiments on the FAU AEC set, we observe obvious improvement of performance by balancing the instance distribution among both classes through random sub-sampling (SIS). Yet, discarding the instances with low agreement levels (AIS) brings a further improvement. A performance comparable with the baseline of the INTERSPEECH 2009 Emotion Challenge is achieved when only 30 % of the whole training set – selected by the proposed method – are used for training. The experiments on the Speaker Likability Database further prove the effectiveness of AIS in the case of discarding training instances.

In future work, the discarding instance number needs to be more discussed when in the blind of test set and type of tasks which are with different distribution of labelling agreement.

5. References

A. Angelova, Y. Abu-Mostafam, and P. Perona. 2005. Pruning training sets for learning of object categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 494–501, San Diego, CA.

A. Angelova. 2004. Data Pruning. M. sci. thesis, California Institute of Tchnology.

F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. 2010. A Database of Age and Gender Annotated Telephone Speech. In *Proc. of LREC*, pages 1562–1565, Valletta, Malta.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

C. E. Erdem, E. Bozkurt, E. Erzin, and A T. Erdem. 2010. Ransac-based training data selection for emotion recognition from spontaneous speech. In *the 3rd international workshop on Affective interaction in natural environments*, pages 9–14, New York, NY.

F. Eyben, M. Wöllmer, and B. Schuller. 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, Florence, Italy.

M. A Fischler and R. C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

D. Fragoudis, D. Meretakis, and S. Likothanassis. 2002. Integrating feature and instance selection for text classification. In *Proc. the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 501–506, Edmonton, Canada.

N. García-Pedrajas, J. Pérez-Rodríguez, and A. de Haro-García. 2013. OligoIS: Scalable Instance Selection for

Class-Imbalanced Data Sets. *IEEE Transactions on Cybernetics*, 43(1):332–346.

M. Grimm and K. Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 381–385, Cancun, Mexico.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

P. Hart. 1968. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3):515–516.

H. Liu and H. Motoda. 2002. On issues of instance selection. *Data Mining and Knowledge Discovery*, 6(2):115–130.

S. Lu, W. Wei, X. Fu, L. Fan, and B. Xu. 2012. Phrase-based data selection for language model adaptation in spoken language translation. In *2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 193–196, Hong Kong, China.

J A. Olvera-López, J A. Carrasco-Ochoa, J F. Martínez-Trinidad, and J. Kittler. 2010. A review of instance selection methods. *Artificial Intelligence Review*, 34(2):133–143.

J. C Riquelme, J. S Aguilar-Ruiz, and M. Toro. 2003. Finding representative patterns with ordered projections. *Pattern Recognition*, 36(4):1009–1018.

G. Ritter, H. Woodruff, S. Lowry, and T. Isenhour. 1975. An algorithm for a selective nearest neighbor decision rule. *IEEE Transactions on Information Theory*, 21(6):665–669.

B. Schuller, S. Steidl, and A. Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proc. of INTERSPEECH*, pages 312–315, Brighton, UK.

B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll. 2011. Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? In *Proc. INTERSPEECH 2011*, pages 1553–1556, Florence, Italy.

B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. V. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. 2012a. The INTERSPEECH 2012 Speaker Trait Challenge. In *Proc. of INTERSPEECH*, Portland, OR. 4 pages.

B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. 2012b. AVEC 2012 – The Continuous Audio/Visual Emotion Challenge. In *Proc. Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012*, pages 449–456, Santa Monica, CA.

B. Schuller. 2013. Multimodal Affect Databases - Collection, Challenges & Chances. In Rafael A. Calvo, Sidney DMello, Jonathan Gratch, and Arvid Kappas, editors, *Handbook of Affective Computing*. Oxford University Press.

I. Sneddon, Ma. McRorie, G. McKeown, and J. Hanratty. 2012. The belfast induced natural emotion database. *IEEE Transaction on Affective Computing*, 3(1):32–41.

S. Steidl. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin.

D R. Wilson and T. R Martinez. 2000. Reduction techniques for instance-based learning algorithms. *Machine learning*, 38(3):257–286.

Y. Wu, R. Zhang, and A. Rudnicky. 2007. Data selection for speech recognition. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 562–565, Kyoto, Japan.

Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller. 2011. Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition. In *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 523–528, Big Island, HY.

Z. Zhang, J. Deng, and B. Schuller. 2013. Co-Training Succeeds in Computational Paralinguistics. In *Proc. 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8505–8509, Vancouver, Canada.