THE OPEN-SET PROBLEM IN ACOUSTIC SCENE CLASSIFICATION

Daniele Battaglino[*†], *Ludovick Lepauloux*[*] *and Nicholas Evans*[†]

[*] NXP Software
Mougins, France

[†] EURECOM
Biot, France

ABSTRACT

Acoustic scene classification (ASC) has attracted growing research interest in recent years. Whereas the previous work has investigated closed-set classification scenarios, the predominant ASC application is open-set in nature. The contributions of the paper are (i) the first investigation of ASC in an open-set scenario, (ii) the formulation of open-set ASC as a detection problem, (iii) a classifier tailored to the open-set scenario and (iv) a new assessment protocol and metric. Experiments show that, despite the challenge of open-set ASC, reliable performance is achieved with the support vector data description classifier for varying levels of openness.

***Index Terms*—** Acoustic scene classification, open-set, support vector data description

1. INTRODUCTION

Acoustic scene classification (ASC) is a research field within the realms of machine-listening and computational auditory scene analysis (CASA) [1, 2, 3]. ASC systems exploit machine learning techniques in order to replicate the human cognitive processes involved in the recognition of ambient sound [4, 5]. Many applications require or can be enhanced with ASC, e.g. context-aware wearable devices, smartphones and robotic systems and a wealth of applications within the so-called Internet of Things (IoT). In these scenarios, intelligent sensing and processing can be applied to optimise or adjust the parameters of a device in sympathy with the immediate environment or use context. An example is the adjustment of a smartphone ring volume when its owner moves from a quiet acoustic environment into a noisier one.

Automatic ASC performance compares favourably to that of human listeners. The work in [6] evaluated a variety of approaches to ASC using a standard, public dataset of 10 classes; the best performing systems achieved a classification accuracy of 75%. More recent work [7], which used similar protocols and metrics and a larger, public dataset comprising 15 hours of audio recordings and 19 different acoustic classes, shows classification accuracies of as high as 91%. Common to all of the past work, is the evaluation of ASC systems in a closed-set scenario for which training data is available for each and every acoustic class which may be encountered during testing.

This evaluation strategy does not reflect practical applications in which out-of-set data may be readily encountered. Without any facility to reject out-of-class acoustic data, its assignment to a target class will result in degraded classification performance. As such, the current closed-set approaches to the evaluation of ASC systems do not reflect the level of performance which could be expected in most practical applications. Surprisingly, to the best of our knowledge, no previous work has investigated ASC in an open-set scenario.

This paper thus reports our attempts to evaluate ASC in an open-set scenario in which the acoustic classes defined for training are a subset of those encountered during testing. We illustrate the limitations of closed-set evaluation, propose a new classifier, protocol and metric for open-set evaluation as a detection problems.

The remainder of the paper is organized as follows: Section 2 describes the difference between closed and open-set scenarios; Section 3 describes different approaches to classification; Section 4 reports experimental work, whereas Section 5 presents our conclusions and some directions for further work.

2. CLOSED VERSUS OPEN-SET

This section describes the difference between closed and open-set scenarios. This necessitates the definition of some notation which is used to describe a measure of openness.

2.1. Closed set

ASC systems are usually developed using large collections of heterogeneous data. The data is aligned to a taxonomy in order to organise the collection into a number of groups or sub-groups which together span the data domain [8]. The groups are referred to as 'classes' or 'contexts' which gather together subsets of data which share similar characteristics. Examples are the classes *car*, *office* and *park*, all of which exhibit their own distinguishable characteristics.

The ASC task then involves the development of a statistical pattern recognition system whose aim is to predict the class to which an unlabelled sample should be assigned. A general approach to ASC thus involves the comparison of data samples to models of each acoustic class. When the universe of classes is exclusively predefined, and thus each sample must necessarily be assigned to one of the classes within, then the task is referred to as being *closed-set*. All existing ACR datasets and evaluations follow such a closed-set paradigm [6].

The notion of a closed universe of classes is perhaps not representative of many practical applications in which the variation in acoustic scene is uncontrolled; a closed-set ASC system developed to recognise *car*, *office* and *park* contexts may fail if it were to encounter *street* noise. It is argued here that most practical applications are indeed uncontrolled and thus ASC solutions must necessarily be able to handle out-of-class data.

2.2. Open set

The ability to reject acoustic data which does not belong, or is not sufficiently close to one of the pre-defined classes is necessary in most practical scenarios. A means of detecting out-of-class data has potential to avoid misclassification and thus to improve system performance beyond what would otherwise be achievable with a closed-set system. Such an open-set system is easily realised with the addi-

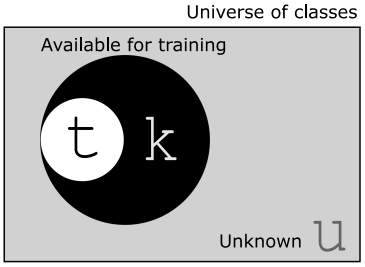


Fig. 1: The universe of acoustic classes. Representative data from target and known negative classes $t \cup k$ are used for training. Representative data from unknown classes is used only in an open-set evaluation. By definition in ASC $t \subseteq k \subseteq u$.

tion of a garbage class to which should be assigned all acoustic data deemed insufficiently close to any of the other defined classes.

Evaluation must then include out-of-class data. Examples for the previously described application could include *street*, *train* or *supermarket* noise. Out-of-class data should be as broad as necessary in order to reflect the practical application. The union of pre-defined and out-of-class data then makes up the entire acoustic universe. While the concept of closed and open-set problems can be clearly defined, the need to evaluate ASC performance in an open-set scenario leads to a *relative* concept of openness. This first requires the definition of some notation.

2.3. The concept of openness

An ASC system is designed to classify a number of *target* classes. In addition to the target classes there is a number of *known* negative classes. Any data sample not in either of these two classes is designated as *unknown*. This arrangement is illustrated in the Venn diagram of Fig. 1. Formally, an open-set evaluation will thus involve some combination of t target classes, k known negative classes and u unknown negative classes. Their values are set according to an evaluation scenario or protocol as follows: a *training* dataset is composed of data from classes t and k while a *testing* dataset combines data from known classes t and k with additional data from unknown classes u.

The need for evaluation and the particular scenario impose some constraints on the values of t, k and u. While u is by its very definition unbounded, the evaluation of ASC systems necessitates the definition of a notionally finite number of unknown classes. The value of all three bears influence on the difficulty of an evaluation; tasks involving greater values of u and k are comparatively more difficult than tasks with smaller values. In particular, unknown negative classes are comparatively more difficult to handle than known negative classes. Related work [9] defines a measure, referred to 'openness', which reflects the difficulty of such a classification task. Drawing upon the original work, the measure of openness is here expressed in terms of t, k and u as:

$$\text{openness} = 1 - \sqrt{\frac{t+k}{t+k+u}} \tag{1}$$

An openness of 0 infers a closed-set problem, while an openness of 1 would infer an entirely open problem. The square root tempers rapid increases in openness with only moderate u. Given a fixed number of targets t, the level of openness depends on k and u: when $u \gg k$, the level of openness will tend to 1; when $u \approx 0$ the level of openness will tend to 0. The relationship between the openness and the number of training classes $t + k$ and testing classes $t + k + u$ is illustrated in Fig. 2.

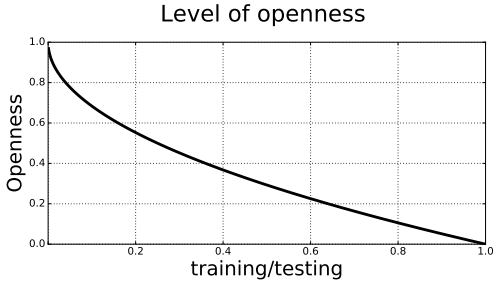


Fig. 2: A plot of openness against the ratio of the number of training classes $(t + k)$ and testing classes $(t + k + u)$ according to Eq. 1. The openness increases as the number of unknown negative classes u increases.

Dataset	t	k	u	openness
DCASE 2013 [6]	10	0	0	0%
Rouen 2015 [7]	19	0	0	0%
DCASE closed-set	1	9	0	0%
Rouen closed-set	1	18	0	0%
DCASE open-set (4 targets)	4	4	2	10%
DCASE open-set (1 target)	1	4	5	29%
Rouen open-set (4 targets)	4	4	11	35%
Rouen open-set (1 target)	1	4	14	48%
Rouen open-set (1 target.)	1	1	14	67%

Table 1: Examples of openness for two well-known datasets, standard closed-set ($u = 0$) and non-standard open-set ($u > 0$) protocols. Openness then varies as a function of the number of target classes t, known negative classes k and unknown negative classes u.

While publicly available datasets for ASC do not preclude an open-set evaluation, standard evaluation protocols are all closed-set ($u = 0$). The second and third rows of Table 1 illustrate the openness of the standard, closed-set evaluation protocols for the DCASE 2013 [6] and Rouen 2015 [7] datasets. Also illustrated in the lowest five rows of Table 1 are different levels of openness for non-standard protocol adaptations which are discussed later.

3. CLASSIFIERS

This section describes two different classifiers. The first is the popular support vector machine (SVM) classifier which is used widely for closed-set ASC. The second is a new approach in the context of ASC and one better suited to open-set classification.

3.1. SVM

Binary classifiers provide a natural solution to ASC. They learn a discrimination function from representative training data from both target and known negative classes. SVM classifiers are one example which project data into a higher-dimensional space in which target and negative data is linearly separable. Separation is obtained with a hyperplane which maximizes the margin between target and negative classes, thereby minimizing classification errors. Previous work in ASC [10, 11] has demonstrated successful results using SVM classifiers in closed-set scenarios. Binary SVM classifiers have also been applied to open-set problems [12]. Even though good separability

can be achieved, generalisation to unknown negative data is typically poor [9]. One principle reason for this is the reliance upon specific negative training data which can never be fully representative of the true variance of negative data in an open-set scenario.

3.2. SVDD

So-called one class SVM approaches have been investigated in the context of many different open-set problems, including image anomaly detection [13], machine fault detection [14] and spoofing detection for speaker verification [15]. One particular approach, referred to as support vector data description (SVDD), learns a hypersphere in which target samples are contained [16]. The goal is to represent target data within the smallest possible hypersphere volume. By using target data only for training purposes, SVDD avoids overfitting to known negatives and thus offers greater generalisation to unknown negatives in an open-set scenario.

The hypersphere has centre a and radius R which are adjusted to contain a percentage of training data X. Based upon the intuition that false positives will be reduced by minimising the volume within the hypersphere, parameters a and R are learned to minimise the so-called structural error:

$$\begin{aligned} \epsilon_{struct}(R,a) &= C\sum_{i}^{N}\xi_i + R^2 \\ s.t. \quad ||x_i - a||^2 &\leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall i \end{aligned} \tag{2}$$

where x_i is the i^{th} training data sample, ξ_i is a slack variable and where C is a penalty factor. The slack variable reflects the distance of the data sample from the hypersphere whereas C reflects the trade-off between hypersphere volume and the percentage of training data contained within.

Lagrangian procedures to optimise a, R, and ξ are described in [16]. Well-known algorithms [17] exist to solve quadratic optimization problems and to find optimal values for the Lagrangian multipliers α_i. The local maximum of the Lagrange function L is found by setting partial derivatives of R, a_i and ξ in Eq. 2 to zero, leading to the following optimization problem and constraints:

$$\begin{aligned} \max L &= \sum_{i}^{N}\alpha_i(x_i \cdot x_i) - \sum_{i,j}^{N}\alpha_i\alpha_j(x_i \cdot x_j) \\ s.t. \quad \sum_{i}^{N}\alpha_i &= 1, \quad a = \sum_{i}^{N}\alpha_i x_i, \quad 0 \leq \alpha_i \leq C, \quad \forall i \end{aligned} \tag{3}$$

The solution to Eq. 3 gives the set of α_i parameters which characterizes the SVDD model. For values of $\alpha_i = 0$, data sample x_i will be within the hypersphere. For $\alpha_i > 0$, x_i will be on the boundary or outside the boundary. Data samples on or outside the boundary are referred to as *support vectors* (SVs). For $C < 1$, some data samples will lie outside the sphere. In this case $\alpha_i = C$ denotes samples outside the hypersphere which are considered as target outliers. Samples for which $0 < \alpha_i < C$ identify support vectors which lie on the boundary. They are referred to as boundary support vectors (BSVs). The radius of the hypersphere is the distance from its centre to one of the BSVs:

$$R^2 = (x_k \cdot x_k) - 2\sum_{i}^{N}\alpha_i(x_i \cdot x_k) + \sum_{i,j}^{N}\alpha_i\alpha_j(x_i \cdot x_j) \tag{4}$$

A data sample lies within the hypersphere if its distance from the centre is less than the radius. Denoting a test sample by z, the distance is thus determined according to:

$$||z - a||^2 = (z \cdot z) - 2\sum_{i}^{N}\alpha_i(z \cdot x_i) + \sum_{i,j}^{N}\alpha_i\alpha_j(x_i \cdot x_j) \tag{5}$$

The decision function $f(z,\alpha)$ is given by $sign(R^2 - ||z-a||^2)$. Finally, data inputs x are mapped into a higher dimensional space where the separability between target and non-target is maximal. As for the regular SVM, the kernel trick avoids the need to operate explicitly in the higher space [18]. The most flexible kernel function in many real-case scenarios, and that used here, is the Gaussian kernel [19, 20].

4. EXPERIMENTAL WORK

This section reports an evaluation of ASC in open and closed-set scenarios. The evaluation is performed in a single-class detection mode. Detection, as opposed to classification, allows for assessment with a comparatively simple metric [21] and also gives a more reliable indication of performance which is less influenced by the number of classes in the dataset. It is stressed, however, that this approach does not preclude multi-class classification which could be implemented straightforwardly with multiple detectors [22].

4.1. Implementation

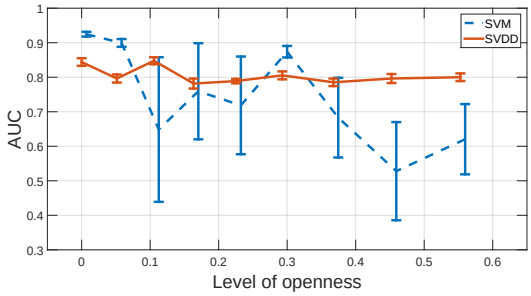
Stereo audio recordings are first converted to mono recordings by channel averaging. Using RASTAMAT tools [23] with default settings, 12 standard Mel frequency cepstral coefficients (MFCCs), without C0, are extracted from windows of 32ms with a window overlap of 16ms. Each audio recording is then represented with the mean and standard deviation of MFCC coefficients thereby producing feature vectors of dimension 24.

SVM and SVDD classifiers are both implemented using the libSVM library [24] using a radial basis function (RBF) kernel and the parameters for this kernel are tuned independently for each classifier. For the SVM classifier, parameter tuning is performed using cross-validation based on the confusion matrix. This approach is not possible for the SVDD classifier which uses knowledge of only the target data. Instead, parameters are tuned independently of known negative data by minimizing the number of SVs while maximizing the radius, as reported in [25]. Finally, feature vectors are normalized according to the *z-score* method [26].

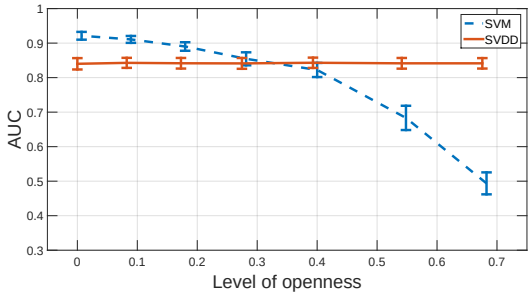
4.2. Datasets and Protocols

The DCASE 2013 [6] and Rouen 2015 [7] datasets are used for evaluation. For DCASE 2013, we used the development set for training and the evaluation set for testing. For Rouen 2015, testing is performed using a *5-fold* cross-validation. In both cases, evaluation involves a gradual transition from closed-set to progressively more and more open-set configurations. Reported first are results for a closed-set evaluation which corresponds to the configurations of the second and third rows of Table 1.

Acoustic class models are learned independently for each target. SVM training is performed using data from both $t = 1$ target class and k known negative classes. In contrast, the SVDD classifier is trained using target class data alone. In order to vary the degree of openness, the number of known negative classes k is varied in



(a)



(b)

Fig. 3: Plots of area under the curve (AUC) against openness for (a) DCASE 2013 and (b) Rouen 2015 datasets for SVM (dashed-blue profiles) and SVDD (solid-red profiles) classifiers. Variance is illustrated with vertical bars.

both cases from 1 to $N-1$, where N is the total number of classes involved in the evaluation ($N = 10$ for DCASE 2013 and $N = 19$ for Rouen 2015).

Testing is performed using varying quantities of data from the whole acoustic universe encompassing t, k and u. When $k = N - 1$, the evaluation is closed-set. The number of unknown acoustic classes in this case is $u = N - t - k = 0$. To better illustrate the closed-set protocol, consider the detection of the *bus* class using the Rouen dataset where $N = 19$. If the number of known negative classes is set to $k = 4$, then the number of unknown negative classes is $u = 14$. According to Eq. 1, this setup corresponds to an openness of 48% as illustrated in the penultimate row of Table 1.

In practice, the performance of the SVM classifier which exploits known negative data will depend on exactly what composes the k known negative classes. Accordingly, in order to marginalise this effect on performance, 10 experiments are performed with different random selections of k known negative classes. Only the average result is reported.

4.3. Metric

Classification accuracy is a popular metric for the evaluation of ASC systems [6]. However, the intrinsic limitations of classification accuracy [27] mean it is ill-suited to open-set problems. Consequently, the area under the curve (AUC) metric is used for all work reported in this paper. The AUC is not influenced by the ratio of target and negative classes and is threshold independent [28]. Scores are extracted using the Platt probability for the SVM classifier [29] and according to the decision function $R^2 - ||z - a||^2$ for the SVDD classifier. The AUC is reported for each classifier and for different levels of openness, averaged over all $t = 1...N$ classes and 10 different compositions of k known negative classes.

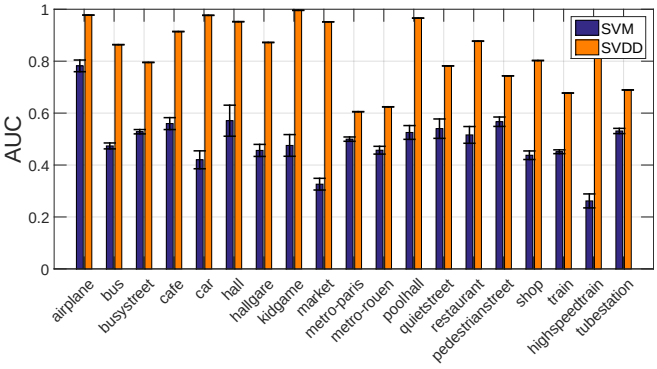


Fig. 4: Individual class AUC results for the SVM and SVDD classifiers for the Rouen dataset with an openness of 0.67.

4.4. Results

Results are illustrated for the DCASE 2013 and the Rouen 2015 datasets in Figs. 3 (a) and (b) respectively. Results for the SVM classifier are illustrated by dashed-blue profiles. Those for the SVDD classifier are illustrated by solid-red profiles. The variance in AUC is also illustrated for each experiment with vertical bars.

Similar trends are observed for both datasets. As the openness increases, the performance of the SVM classifier deteriorates, falling from 95% to 60% for the DCASE 2013 dataset and from 90% to 50% for the Rouen 2015 dataset. In contrast, results for the SVDD classifier remain relatively stable for both datasets, measuring in the order of 80% and 85% for the DCASE 2013 and Rouen 2015 datasets respectively.

Fig. 4 illustrates separately the AUC for each class in the Rouen 2015 dataset for a openness of 0.67. Consistent with results illustrated in Fig. 3, the SVDD classifier outperforms the SVM classifier. Of greater interest here, however, is the variation in performance for different compositions of k known negative classes, again illustrated in terms of variance with vertical bars. While the performance of the SVM classifier is impacted by the specific combination of k known negative classes, that of the SVDD classifier is relatively unaffected. The significance of the two methods is measured with a McNemar test[30] which rejects the hypothesis at 95% significance that the two classifiers have equal predictions.

5. CONCLUSIONS

This paper reports the first attempt to develop an approach to acoustic scene classification (ASC) in a practical, open-set scenario. A traditional ASC classifier is shown to outperform an open-set classifier in a largely closed scenario. When the level of openness increases, however, performance degrades rapidly, whereas the performance of the newly proposed approach to open-set ASC remains stable. The support vector data description (SVDD) classifier learns a hypersphere from target data only. While using target data only for training, this classifier is less susceptible to over fitting to known negative data and is thus more reliable in the face of unknown negative data. The paper also introduces a new approach to assessment based on a detection formulation, a new protocol and metric. Given that the predominant ASC use-case scenario is open-set in nature, it is hoped that the approach to assessment reported in this paper is adopted by the research community for further work.

References

[1] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[2] D. P. W. Ellis. "Prediction-Driven Computational Auditory Scene Analysis". PhD thesis. Cambridge MA: MIT, Dept. of Electrical Engineering and Computer Science, 1996.

[3] H. G. Okuno, T. Nakatani, and T. Kawabata. "Cocktail-Party Effect with Computational Auditory Scene AnalysisPreliminary Report". In: *Advances in Human Factors/Ergonomics* 20 (1995), pp. 503–508.

[4] M. Stephen. *Recognition of sound sources and events*. Thinking in sound: the cognitive psychology of human audition. London: Oxford Univ. Press, 1993.

[5] A. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. A Bradford book. Bradford Books, 1994.

[6] D. Barchiesi et al. "Acoustic Scene Classification: Classifying environments from the sounds they produce". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 16–34.

[7] A. Rakotomamonjy and G. Gasso. "Histogram of gradients of Time-Frequency Representations for Audio scene detection". In: *arXiv:1508.04909 [cs]* (Aug. 2015). arXiv: 1508.04909.

[8] L. Lu, H.-J. Zhang, and H. Jiang. "Content analysis for audio classification and segmentation". In: *IEEE Transactions on Speech and Audio Processing* 10.7 (2002), pp. 504–516.

[9] W. J. Scheirer et al. "Towards Open Set Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 36 (7 2013).

[10] G. Roma et al. "Recurrence quantification analysis features for auditory scene classification". In: *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events* (2013).

[11] D. Battaglino et al. "Acoustic context recognition using local binary pattern codebooks". In: *WASPAA 2015, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 18-21 October 2015, New Paltz, NY, USA*. New Paltz, UNITED STATES, Oct. 2015.

[12] Q. Zhao and J. C. Principe. "Support vector machines for SAR automatic target recognition". In: *IEEE Transactions on Aerospace and Electronic Systems* 37.2 (2001), pp. 643–654.

[13] A. Banerjee, P. Burlina, and C. Diehl. "A support vector method for anomaly detection in hyperspectral imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 44.8 (2006), p. 2282.

[14] A. Ypma, D. Tax, and R. Duin. "Robust machine fault detection with independent component analysis and support vector data description". In: *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.* Aug. 1999, pp. 67–76.

[15] F. Alegre, A. Amehraye, and N. Evans. "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns". In: *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 1–8.

[16] D. M. J. Tax and R. P. W. Duin. "Data domain description using support vectors". In: *Proceedings of the European Symposium on Artificial Neural Networks*. 1999, pp. 251–256.

[17] J. Platt. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Tech. rep. MSR-TR-98-14. Microsoft Research, 1998, p. 21.

[18] B. Schiilkopf. "The kernel trick for distances". In: *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*. Vol. 13. MIT Press. 2001, p. 301.

[19] S. Wang et al. "A modified support vector data description based novelty detection approach for machinery components". en. In: *Applied Soft Computing* 13.2 (Feb. 2013), pp. 1193–1205.

[20] L. Zhuang and H. Dai. "Parameter optimization of kernel-based one-class classifier on imbalance learning". In: *Journal of Computers* 1.7 (2006), pp. 32–40.

[21] F. Li and H. Wechsler. "Open set face recognition using transduction". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27.11 (2005), pp. 1686–1697.

[22] A. Rabaoui et al. "One-class SVMs challenges in audio detection and classification applications". In: *EURASIP Journal on Advances in Signal Processing* 2008.1 (2008), pp. 1–14.

[23] D. P. W. Ellis. *PLP and RASTA (and MFCC, and inversion) in Matlab*. Software available at `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`. 2005.

[24] C.-C. Chang and C.-J. Lin. "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 27:1–27:27.

[25] A. Theissler and I. Dear. "Autonomously determining the parameters for SVDD with RBF kernel from a one-class training set". In: *Proceedings of the WASET International Conference on Machine Intelligence*. 2013, pp. 1135–1143.

[26] I. B. Mohamad and D. Usman. "Standardization and its effects on k-means clustering algorithm". In: *Res. J. Appl. Sci. Eng. Technol* 6.17 (2013), pp. 3299–3303.

[27] F. J. Valverde-Albacete and C. Pelez-Moreno. "100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox". In: *PLoS ONE* 9.1 (Jan. 2014), pp. 1–10.

[28] M. Sokolova and G. Lapalme. "A systematic analysis of performance measures for classification tasks". en. In: *Information Processing & Management* 45.4 (July 2009), pp. 427–437.

[29] J. C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in large margin classifiers*. MIT Press, 1999, pp. 61–74.

[30] M. W. Fagerland, S. Lydersen, and P. Laake. "The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional". In: *BMC Medical Research Methodology* 13.1 (2013), pp. 1–8.