First Experiments to Detect Anomaly Using Personality Traits vs. Prosodic Features

Cedric Fayet[1,2], Arnaud Delhay[1,2]✉, Damien Lolive[1,2], and Pierre-François Marteau[1,3]

[1] IRISA - EXPRESSION Team, Lannion&Vannes, France
[2] Université de Rennes 1, Rennes, France
[3] Université de Bretagne Sud, Vannes, France
{cedric.fayet,arnaud.delhay,damien.lolive,pierre-francois.marteau}@irisa.fr

Abstract. This paper presents the design of an anomaly detector based on three different sets of features, one corresponding to some prosodic descriptors and two extracted from Big Five traits. Big Five traits correspond to a simple but efficient representation of a human personality. They are extracted from a manual annotation while prosodic features are extracted directly from the speech signal. We evaluate two different anomaly detection methods: One-Class SVM (OC-SVM) and iForest, each one combined with a threshold classification to decide the "normality" of a sample. The different combinations of models and feature sets are evaluated on the SSPNET-Personality corpus which has already been used in several experiments, including a previous work on separating two types of personality profiles in a supervised way. In this work, we propose the above mentioned unsupervised methods, and discuss their performance, to detect particular audio-clips produced by a speaker with an abnormal personality. Results show that using automatically extracted prosodic features competes with the Big Five traits. In our case, OC-SVM seems to get better results than iForest.

Keywords: Anomaly detection, Isolation Forest, Isolation Tree, One Class – Support Vector Machine, Threshold Classification, Social Signal, Big Five, Prosody, SSPNET-Personality

1 Introduction

According to Chandola *et al.* [6], "an anomaly is defined as a pattern that does not conform to an expected normal behavior". The main objective of any anomaly detection system is to identify abnormal states from normal state distributions. The feature sets that describe these states are related to the nature of the input data (continuous, categorical, spatial, or spatio-temporal data), but also to the nature of the anomalies that we aim to track (point-wise, contextual or collective anomalies). Finally, to choose an adequate anomaly detection model, we also need to consider the desired output (score or label) and the availability of labeled data (supervised or unsupervised techniques).

In the literature, anomaly detection techniques are applied to intrusion detection [3], fraud detection [1], sensor network [15], monitoring flight safety [12], *etc.* As far as we know, the use of anomaly detection with the speech signal is more focused on speech pathology or disorder [2], or on the deduction of another type of pathology or disease (e.g. cancer) [7]. The speech signal is also used for detecting stress or depression [10, 20], that could be seen as an anomaly detection in the way that it can be considered as an abnormal mental state.

Handling audio, video or biological signal to infer social information such as personality is part of a field called *Personality Computing* [21].

The Big Five model has been proposed to describe the speaker personality through the five following personality traits: openness, conscientiousness, extraversion, agreeableness and neuroticism [11].

Different tests with adaptation to a local context have been conducted on different languages and cultures, and the big five model seems to be generalizable to them [8, 9]. SSPNET-Personality corpus was built to experiment the prediction of the big five scale over audio features [14]. It is composed of French audio clips extracted from the ”Radio Suisse Romande”. These clips are records of professional (journalists) and guest speakers.

In [19], Schuller *et al.* offers a good overview of the different systems using this corpus. A lot of experiments have already been conducted on this corpus and most of them try to predict the value of the big five scales using different audio representations. In addition, some experiments done by Mohammadi and Vinciarelli [14] suggest that the big five representation could be a good predictor of the role of a speaker. The authors propose a supervised SVM method that learns to recognize if a speaker is a professional one or a guest. The feature set used in this work is the psychological evaluation of a sample given by annotators using Big five traits. Their predictor reaches an accuracy of about 75%.

The purpose of our work is to compare the use of the big five features to the use of prosodic features to design a personality predictor able to separate professional speakers from non-professional guests. To this end, we consider a professional speaker as belonging to a normal personality class and a guest to an abnormal one. Contrary to [14], we propose to work in the unsupervised anomaly detection framework and we propose to evaluate two unsupervised strategies. Three sets of features are evaluated. Results show that prosodic features perform well and are less costly compared to manually annotated Big Five traits, even if those last features may provide better detection accuracy.

The remainder of the paper is structured as follows. The proposed anomaly detection method is described in section 2. In section 3, the feature sets that we compare are detailed before describing the experimental setup in section 4. Finally, the results are presented and discussed in section 5.

2 Method

This work aims at determining if a sample is normal (professional speaker) or abnormal (guest speaker) based on a train set containing only normal audio clips.

Each audio clip is described by a feature vector summarizing the time evolution of the features as described in section 3.

In this section, we describe the two methods that we have used to perform the anomaly detection: an Isolation Forest approach (referred to as iForest in the paper) and a One-Class Support Vector Machine approach (referred to as OC-SVM in the paper). The OC-SVM has been used to ensure a comparison baseline with the SVM method used in [14].

2.1 Isolation Forest (iForest)

Isolation Forest is an ensemble learning method designed to detect anomaly. The particularity of this method is that it explicitly isolates anomalies rather than learns a model for normal instances [13]. The main assumption behind iForest is that a normal sample is hard to isolate from other samples, and on the contrary, an anomaly is more easily isolated from other samples. iForest is composed of T iTrees, each one built on a random selection of ψ samples drawn from the training set. From this subset of samples, an iTree is constructed by a random recursive partitioning, until all the samples are isolated or until a stop criterion is reached (a depth limit for example). The partitioning is realized by the random selection of an attribute (feature) and the random choice of a pivot value in the range of the selected attribute. For an iTree, the sample score is computed as the path length between the leaf node containing the sample and the root node of the tree.

Let x be a sample, n the number of samples on which the iTrees are built. Let f be the iForest, with $f = \{t_1, t_2, ...t_T\}$. Let $h(t, x)$ be the number of edges of the t iTree between the root and the leaf which contains (or isolates) x. Let $c(n)$ be the average path length of unsuccessful search in a Binary Search Tree. $c(n)$ estimates the average path length of an iTree.

The anomaly score s of an instance x estimated with the f iForest is given equation 1.

$$s(\mathbf{x}, \mathbf{f}, \mathbf{n}) = 2^{\frac{-\sum_{k=1}^{T} h(\mathbf{f_k}, \mathbf{x})}{T * c(\mathbf{n})}} \tag{1}$$

2.2 One-Class SVM Approach (OC-SVM)

In short, a SVM classifier learns a boundary which maximizes the margin between classes. This well-known approach has been shown to be very effective on many classification problems. OC-SVM is an adaptation of SVM to the one-class problem. After transforming the feature space via a kernel, OC-SVM considers as a starting point, all the available data as member of a single class $C_{inliers}$ and the origin in the space defined by the nonlinear kernel as the only member of a class $C_{outliers}$ [18].

During the training, the hyper-parameter ν corresponds to a penalizing term which represents a trade-off between inliers and outliers. With the SVM approaches, the choice of the kernel is important to improve the results. The most widely used kernels are linear (inner product), polynomial, RBF and sigmoid. Related to the kernel used (RBF, polynomial, sigmoid) we have to determine its parametric coefficients. For instance, with the sigmoid kernel :

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x}^T \mathbf{y} + c_0)$$

where $\mathbf{x}$ and $\mathbf{y}$ are input vectors and γ (slope) and c_0 (intercept) are parametric values related to the kernel.

We choose to use the distance between a new sample and the learned boundary as an anomaly score, instead of directly deciding which class the samples belong to. We then use a classification threshold based on this anomaly score to decide if a sample is normal or abnormal. This two steps strategy introduces some adjustable fuzziness around the boundary.

3 Materials

The experiments are conducted over the SSPNET-Personality Corpus [14]. The corpus contains 640 audio clips divided into 307 audio clips of professional speakers and 333 audio clips of guests in French language. The duration of each audio clip is about 10 seconds, based on the assumption that it takes short time to get an opinion about others personality.

3.1 Big Five Features

For each sample, 11 non-native French speakers evaluate the BFI-10 Questionnaire [17] from which a score is computed for each Big Five's scale. Consequently, 11 evaluations of the big five features are available. For our experiment, we consider two sets of features based on the big five model:

- *BigFive5*: for each sample and for each Big Five scale, we compute the mean of the 11 evaluations which leads to 5 features.
- *BigFive55*: for each sample, we concatenate the 11 evaluations given for each sample which leads to 55 features.

The purpose of this last set of features is to verify if the information contained in several distinct annotations is complementary or can simply be aggregated in a lower dimension feature vector, as in *BigFive5* feature vector. The main drawback of these two feature sets is that they are the result of manual annotation, which is furthermore difficult to predict from the speech signal only.

3.2 Prosodic Features

Prosodic features are commonly used to capture affect cues in a speech signal. Contrary to the Big Five traits, a large number of prosodic features are much easier to extract automatically from the speech signal. In this study, we adopt the 6 dimensional prosodic feature set as described in [14]. Using the PRAAT software [4], we extract the pitch, the first two formants, the energy and the duration of voiced/unvoiced segments with a sliding analysis window size of 40 ms with a step of 10 ms. From these low-level features, resulting from the extraction, we derive the final features that summarize their time evolution by computing mean, maximum, minimum and entropy values for each of the 6 features. Consequently, the final *Prosodic* feature set is composed with 24 features for each audio clip.

3.3 Pre-processing

Normalizing the features is an important preprocessing step before using the OC-SVM approach. Features that do not share the same range of values and the same variations could affect the quality of the OC-SVM model. Therefore this step is really important in the case of *Prosodic* features, which are composed of different types of features. We choose to perform a standardization (zero mean and unit variance) on all types of features.

4 Experimental Setup

The different experiments described in this section are carried out by following the same procedure and use python 3.5 and scikit-learn [16]. From the corpus, we build three sets (train, test, and validation sets), as described below. To increase the statistical confidence of our results, we run each experiment 60 times by distributing randomly the samples on the three sets. For each experiment, we compute the mean and the standard-deviation divided by the mean of the different runs. The data is thus split into three folds as follows:

- Train set: 207 clips of professional speakers and a variation from 0 to 103 guest clips.
- Test set: 50 clips of professional speakers and 50 clips of guests.
- Validation set: 50 clips of professional speakers and a variation from 0 to 50 of guest clips.

4.1 Hyper-parameters Tuning

According to the available data, since 207 samples are available in the training set, we need to be careful about the number of parameters in our model to avoid over-fitting.

In the case of the OC-SVM approach, after testing different types of kernel in our experiments, we have chosen to present the results with the sigmoid kernel which gives the best results (on the train data). With the sigmoid kernel, we need to determine three hyper-parameters ν, γ and c_0. For the rest of the paper, we choose to fix c_0 at 0. Without considering the fuzzy boundary, the OC-SVM gives a first classification of the sample. By using it, we compute a classification score [5] which can be used to evaluate the classification quality and consecutively the

Table 1. Hyper-parameters chosen for each AD

	iForest		OC-SVM	
	T	ψ	γ	ν
BigFive5	17	0.29	0.016	0.7
BigFive55	77	0.31	0.007	0.6
Prosodic	59	0.26	0.009	0.57

quality of the hyper-parameters. An exhaustive grid search is used to determine the values of the two hyper-parameters. For ν, we considered a range between 0.001 and 1 with a step of 0.025. For γ, the range is between 0.0001 and 1 with a step of 0.0001. The classification score is used to elect the best couple of hyper-parameters.

In the case of the iForest approach, we need to find two hyper-parameters: ψ (sub-sampling size given as a percentage of the available data) and t (number of sub-estimators *i.e* iTrees). When using an iForest, each sample is associated to an anomaly score. We consider 10 percents of the samples as abnormal samples, those with the highest probability to be abnormal. With this assumption, we can use a classification score as we did previously, and then, use an exhaustive grid search to get the two hyper-parameters. The range for ψ is between 0.1 and 0.9 with a step of 0.01, and for t between 10 and 200 with a step of 10.

For each set of features, we considered a training set with all the available normal samples to determine the hyper-parameters (table 1) with respect to the aforementioned methods.

4.2 Experiments

To compare the different sets of features according to the chosen approach, we carry out the following steps: for each set of values (*BigFive5*, *BigFive55*, *Prosodic*), we test each anomaly detection model after tuning hyper-parameters as explained before (section 4.1). Then, we compare the different models by using a ROC (Receiver Operating Characteristic) curve. It means that for a given FPR (False Positive Rate), we search the associated TPR (True Positive Rate). This sampled association is obtained by testing a range of possible threshold values. The ROC curve gives an information about the detector quality as a function of the threshold value.

The second step consists in estimating the robustness of each detector to a degradation of the training set. In this purpose, we introduce a certain percentage of abnormal samples into the training set (0% to 50%). We keep the different hyper-parameter values unchanged.

The last step consists in observing anomaly scores obtained for the different sets of features with both approaches. Considering the hyper-parameters obtained previously, we train the predictor with all the available normal samples. Then, we compute anomaly scores for normal and abnormal samples separately.

Table 2. Mean area under ROC curve (ROC-AUC score) and standard deviations, for the two approaches combined with the three feature sets.

	BigFive5	BigFive55	Prosodic
OC-SVM	0.857±0.030	0.918±0.024	0.876±0.036
iForest	0.762±0.031	0.890±0.032	0.801±0.037

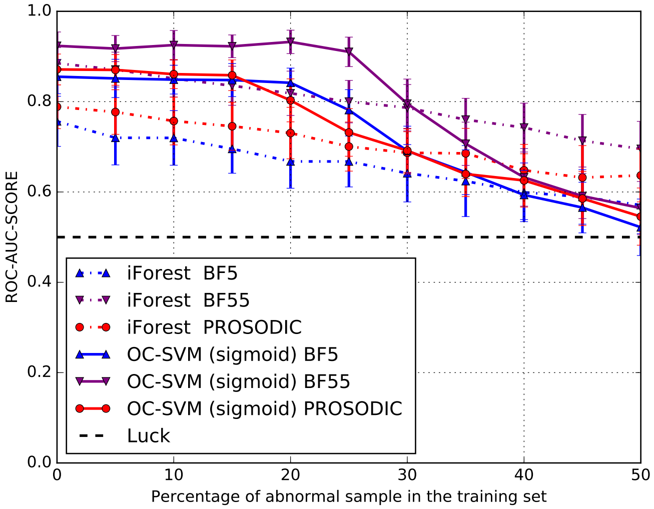

Fig. 1. Models Robustness to the introduction of abnormal samples in the training set.

5 Results

Our first experiment step (table 2) consists in evaluating the quality of each feature set. With our data, the *BigFive55* feature vector, regardless of the method used, performs better than the others. If we consider the OC-SVM approach, the *Prosodic* and *BigFive5* feature vectors get comparable results. For the iForest approach, *BigFive5* feature vector achieves a worst performance than the others. Moreover, the *BigFive5* feature vector has the lowest results for both methods, thus showing that the aggregation of the individual annotations induces a significant information loss.

Our second experiment step (figure 1) consists in evaluating the quality of each AD model when the training set is degraded, *i.e.* by including a certain percentage of abnormal samples in it. The results show that the iForest and OC-SVM approaches have different responses to degradation. The OC-SVM approaches have a degradation in two steps: before a certain percentage of contamination, they are robust to degradation and after this limit their scores decrease down to around 0.5 (random detector) for 50% of degradation. The iForest approaches have a kind of linear degradation along the degradation axis. For the OC-SVM approach, the *BigFive55* feature vector seems to keep stable results for

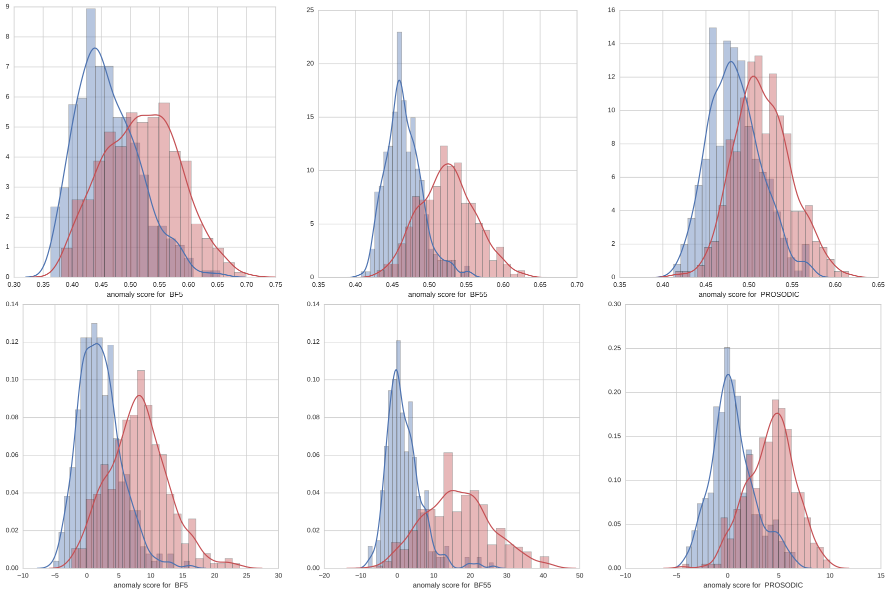

Fig. 2. Anomaly scores for normal samples (blue curve, left curve) and for abnormal samples (red curve, right curve) obtained for each feature set BigFive5 (left), BigFive55 (middle), Prosodic (right) with the two approaches iForest (top) and OC-SVM (bottom).

approximately 25 percents of degradation. The other two feature vectors start to have a decrease of their quality for less than 15 percent of degradation. For the iForest approach, the *Prosodic* feature vector seems to be more robust to degradation than the *BigFive5*. In any case, the *BigFive55* feature vector gives a more robust separation between normal and abnormal samples.

In our final experiment step (fig 2), we analyze the anomaly score obtained for all the available samples in the corpus. We draw one curve for normal samples and another one for abnormal samples. For both approaches, the intersection area between abnormal and normal curves is greater for *BigFive5* than for *Prosodic*, and greater for *Prosodic* than for *BigFive55*. By comparing the curves for *BigFive5* and *BigFive55* with both approaches, we notice that the distributions of anomaly scores for abnormal and normal samples are more different for *BigFive55* than for *BigFive5*. Indeed, by aggregating the annotators of *BigFive55* to get *BigFive5*, we diminish the ability to discriminate between normal and abnormal samples. It is noticeable that the iForest achieves a lower performance than the OC-SVM. It suggests that, in our context, modeling the normality class is better than considering anomalies as isolated points.

6 Conclusion

The main objective of this paper was to compare the use of prosodic cues and the Big Five annotation traits as feature sets for anomaly detection. We have conducted some experiments with the SSPNET-Personality Corpus using professional speakers as normal samples and guest as abnormal samples. This choice was motivated by Mohammadi *et al.* work [14] that demonstrates the effectiveness of using the Big Five features to train a supervised classifier able to separate these two categories of speakers. We built three sets of features (*BigFive5*, *BigFive55* and *Prosodic*) based on the speech signal and a psychological evaluation (Big Five model) available on the data set. We have used two different unsupervised machine learning methods (iForest, OC-SVM) to build our anomaly predictors. Based on the results, the good performance of the *BigFive55* feature set compared to the *Prosodic* feature set indicates that features based on psychological information can bring more information than audio features only. However, the prosodic features are easy to extract from the speech signal and thus seem to be the best compromise between ease of extraction and performance. The better results obtained with OC-SVM compared to iForest seem to indicate that an anomaly in our context is more related to learning a single specific cluster in the data than searching for samples which are really different from the others. A natural follow-up is to test other feature sets: for instance one can increase the number of features to reach a size similar to the *BigFive55* feature vector. Finally, conducting these experiments on other audio corpora, or trying to generalize our results on multimedia data are part of our future work.

Acknowledgments. This research has been financially supported by the French Ministry of Defense - Direction Générale de l'Armement and the région Bretagne (ARED) under the MAVOFA project.

References

1. Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: A survey. Journal of Network and Computer Applications 68, 90–113 (2016)
2. Alonso, J.B., Díaz-de María, F., Travieso, C.M., Ferrer, M.A.: Using nonlinear features for voice disorder detection. In: ISCA tutorial and research workshop (ITRW) on non-linear speech processing (2005)
3. Axelsson, S.: Intrusion detection systems: A survey and taxonomy. Tech. rep., Chalmers University of Technology, Göteborg, Sweden (2000)
4. Boersma, P., Weenink, D.: Praat: doing phonetics by computer, http://www.praat.org/
5. Caliski, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3(1), 1–27 (1974)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) 41(3), 15 (2009)
7. Clapham, R.P., Molen, L.v.d., Son, R.J.J.H.v., Brekel, M.W.M.v.d., Hilgers, F.J.M.: NKI-CCRT corpus - speech intelligibility before and after advanced head

and neck cancer treated with concomitant chemoradiotherapy. In: Proc. of the 8th international conference on Language Resources and Evaluation (LREC) (2012)

8. Goldberg, L.R.: Language and individual differences: The search for universals in personality lexicons. Review of personality and social psychology 2(1), 141–165 (1981)
9. Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., Lero Vie, M.: How universal is the Big Five ? Testing the five-factor model of personality variation among foragerfarmers in the bolivian amazon. Journal of Personality and Social Psychology 104(2), 354–370 (2013)
10. He, L., Lech, M., Maddage, N.C., Allen, N.: Stress detection using speech spectrograms and sigma-pi neuron units. In: Fifth International Conference on Natural Computation ICNC'09. vol. 2, pp. 260–264. IEEE (2009)
11. John, O.P., Srivastava, S.: The Big Five trait taxonomy: History, measurement, and theoretical perspectives, vol. 2, pp. 102–138. Guilford (1999)
12. Li, L., Gariel, M., Hansman, R.J., Palacios, R.: Anomaly detection in onboard-recorded flight data using cluster analysis. In: IEEE/AIAA 30th Digital Avionics Systems Conference. pp. 4A4–1–4A4–11. IEEE (2011)
13. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: Eighth IEEE International Conference on Data Mining. pp. 413–422 (2008)
14. Mohammadi, G., Vinciarelli, A.: Automatic personality perception: Prediction of trait attribution based on prosodic features. IEEE Transactions on Affective Computing 3(3), 273–284 (2012)
15. Park, K., Lin, Y., Metsis, V., Le, Z., Makedon, F.: Abnormal human behavioral pattern detection in assisted living environments. In: Proc. of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments. p. 9 (2010)
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830 (2011)
17. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. Journal of Research in Personality 41(1), 203–212 (2007)
18. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural computation 13(7), 1443–1471 (2001)
19. Schuller, B., Steidl, S., Batliner, A., Nth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. Computer Speech & Language 29(1), 100–131 (2015)
20. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proc. of the 6th International Workshop on Audio/Visual Emotion Challenge. pp. 3–10. ACM (2016)
21. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Transactions on Affective Computing 5(3), 273–291 (2014)