MDPI

Article

Gender recognition by voice using an improved self-labeled algorithm

Ioannis E. Livieris[1,*], Emmanuel Pintelas[2*], Panagiotis Pintelas[3]

[1] Department of Computer & Informatics Engineering, Technological Educational Institute of Western Greece, GR 263-34, Greece.
[2] Department of Electrical & Computer Engineering, University of Patras, GR 265-00, Greece; ece6835@upnet.gr.
[3] Department of Mathematics, University of Patras, GR 265-00, Greece; ppintelas@gmail.com
* Correspondence: livieris@teiwest.gr

Version March 2, 2019 submitted to Machine Learning and Knowledge Extraction

Abstract: Speech recognition has various applications such as human to machine interaction, sorting of telephone calls by gender categorization, video categorization with tagging and so on. Nowadays, machine learning is a popular trend which has been widely utilized on various fields and applications, exploiting the recent development in digital technologies and the advantage of storage capabilities from electronic media. Recently, research focuses on the combination of ensemble learning techniques with the semi-supervised learning framework aiming to build more accurate classifiers. In this paper, we focus on gender recognition by voice utilizing a new ensemble semi-supervised self-labeled algorithm. Our preliminary numerical experiments demonstrate the classification efficiency of the proposed algorithm in terms of accuracy, leading to the development of stable and robust predictive models.

Keywords: Semi-supervised learning; self-labeled methods; ensemble learning; gender recognition; classification.

1. Introduction

Speech constitutes one of the most popular and significant means for humans to communicate, express their emotions, cognitive states, and intentions to each other. The speech is produced by humans with a natural biological mechanism in which lungs discharge the air and convert it to speech passing through the vocal cords and organs like tongue, teeth, lips etc. [1]. In general, a speech and voice recognition system can be used for gender identification. A natural voice recognition system is the human ear. Human ear has an excellent mechanism which can efficiently distinguish the gender by voice and speech based on attributes like frequency and loudness. In a similar way, a machine can be taught to do the same thing by choosing and incorporating the right features from voice data on a machine learning algorithm.

Gender recognition is a technique which is often utilized to determine the gender category of a speaker by processing speech signals. Speech signals taken from a recorded speech can be used to acquire acoustic attributes such as duration, intensity, frequency and filtering [2]. Some applications that gender recognition can be useful are: speech emotion recognition, human to machine interaction, sorting of telephone calls by gender categorization, automatic salutations, muting sounds for a gender and audio/video categorization with tagging (see [1,3–5] and the references therein).

As technology is growing in a rapid way, machine learning is a research field which has met a major development in our days; thus it has been widely established as a popular trend. Machine learning is a subset of artificial intelligence which utilizes algorithms and data to teach computers

make decisions on specific problems in various fields like finance, banking, medicine etc. [6–8] Along this line, several studies for gender recognition and identification by voice using machine learning and data mining techniques have been conducted [1,9–11]. However, the development of an accurate prediction model for gender recognition by voice is still considered a rather difficult and challenging task. In [12,13] conducted an extensive experimental analysis and pointed out the difficulties of this classification problem since speech signals are highly time-varying and have very big randomness. This is mainly due to the fact that the progress in the field has been hampered by the lack of available labeled data for efficiently training a supervised classifier. Furthermore, in order to train efficiently a classifier and be able to make accurate predictions, it often needs a large amount of labeled data. Nevertheless, the process of finding sufficient labeled data for training classifiers to make accurate predictions is often an expensive and time-consuming task as it requires human efforts, while in contrast finding unlabeled data in general is significantly easier. To address the problem of insufficient labeled data, Semi- Supervised Learning (SSL) algorithms constitute the appropriate methodology to exploit the hidden information found in the unlabeled set aiming to build more accurate classifiers. In the literature, several classes of SSL algorithms have been proposed and evaluated, each of them based on different methodologies and techniques related to the link between the distribution of labeled and unlabeled data (see [14–16] and the references there in). Self-labeled algorithms probably constitute the most popular and widely utilized class of SSL algorithms. The algorithms of this class follow an iterative procedure, augment an initial labeled dataset using their own predictions from a large pool of unlabeled dataset. Triguero et al. [14] proposed an in-depth taxonomy of self-labeled algorithms based on their main characteristics and made an exhaustive research of their classification efficacy on various datasets.

Ensemble Learning (EL) is another way of obtaining better results for a higher classification accuracy, which has been developed in the last decades. The main object of this methodology is the combination of several prediction models, in order to build a more accurate model rather than using a single one. Furthermore, the development of algorithms which hybridize SSL and EL approaches is another recent methodology which can be beneficial to each other and can build more robust classification algorithms, leading to even better classification results [17,18].

In this work, we propose a new ensemble-based self-labeled algorithm, called iCST-Voting, for gender recognition by voice. This algorithm combines the individual predictions of three of the most popular and efficient self-labeled methods i.e. Self-training, Co-training and Tri-training utilizing an ensemble as base learner. Our experimental results reveal the efficiency of this algorithm compared against state-of-the-art self-labeled algorithms.

The remainder of this paper is organized as follows: Section 2 presents a synopsis of related work on gender recognition by voice. Section 3 presents a brief description of the self-labeled algorithms. Section 4 presents the proposed classification algorithm. Section 5 presents the datasets and our experimental results. Finally, Section 6 presents our concluding remarks and some directions for future research.

2. Related work

During the last decades, machine learning models and data mining techniques have been widely utilized for gender recognition by voice. These prediction models can identify the gender of a person by utilizing various features such as length of the vocal folds, gait and speech. More specifically, the acoustic properties acquired from voice and speech signals like duration, intensity and frequency can be used as features to recognize the gender of speaker. A number of studies have been carried out in recent years; some useful outcomes of them are briefly presented below.

Maka et al. [10] used 630 speakers, 438 males and 192 females in their experiments for the gender identification problem in different acoustical environments (indoor and outdoor auditory scenes). In addition, for the evaluation stage each sentence has been mixed with several types of background

noise. In their results, they found out that non-linear smoothing increases the classification accuracy by 2% and the recognition accuracy obtained was 99.4%.

Bisio et al. [4] developed an Android SPEech proCessing plaTform as smaRtphone Application (SPECTRA) for gender, speaker and language recognition by utilizing multiple unsupervised support vector machine classifiers. An interesting and innovative point in this work is the dynamic training with the features extracted from every user, having SPECTRA installed on his personal android smartphone. This can lead on building more robust classifiers with higher classification accuracy, resulting in better recognition performances.

Pahwa et al. [1], proposed a recognition system to determine the gender using speech samples of 46 speakers. In particular, they extracted one of the most dominant and most researched speech feature, Mel coefficients and its first and second order derivatives. Their proposed model consists of a support vector machine and neural network classifier using a stacking methodology. The classification accuracy obtained from their numerical experiments was 93.48%.

Pribil et al. [19], proposed a two level GMM algorithm to recognize age and gender. Their proposed classifier was first verified for detection of four age categories (child, young, adult, senior) and for recognizing the gender for all but children's voices in Czech and Slovak languages. The prediction accuracy on gender identification was above 90%. In a similar work, Pribil et al. [20] developed a two level GMM classifier to detect age and gender. The classification accuracy achieved on gender recognition was 97.5%. Furthermore, the obtained gender and age classification accuracy results were compared with the results achieved by the conventional listening test which is an evaluation method of the quality of the synthetic speech [21].

Buyukyilmaz et al. [9], utilized a multilayer perceptron deep learning model using the acoustic properties of the voices and speech to identify the voice gender. The dataset they utilized for their experiments consisted of 3168 recorded samples of human voices. Their classification model managed to achieve 96.74% accuracy. Additionally, they have designed a web page to detect the gender of voice by utilizing the obtained model.

Zvarevashe et al. [11], proposed a gender voice recognition technique utilizing feature selection through the Random Forest recursive feature elimination with Gradient Boosting Machines (GBMs) algorithm for gender classification. Acoustic features were collected from a public gender voice dataset including 1584 males and 1584 females. The GBMs algorithm had obtained an accuracy of 97.58% without feature selection while by applying feature selection it almost achieved 100%.

3. On semi-supervised self-labeled classification algorithms

In this section, we present a short description of the most popular self-labeled classification algorithms proposed in the literature.

Suppose that $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$ is a set of instances, where each example $x \in \mathbb{R}^n$ is a n-dimensional vector of n features belonging to a class y and $(x^{(i)}, y^{(i)})$ is the i-th instance of m total instances. Then, let assume that a training set $L \cup U$ is composed of a labeled set L of N_L instances where y is known and of an unlabeled set U of N_U instances where y is unknown with $N_U >> N_L$ and also a test set T that consists of N_T unseen instances where y is unknown. In general, the basic objective of self-labeled algorithms if to classify unlabeled data based on the most confident predictions in order to augment the initial training set.

The self-labeled methods are divided into two main categories, Self-training [22] and Co-training [23]. In *Self-training*, one classifier is initially trained on a labeled training set and it is used to make predictions for the examples of an unlabeled set, where the most confident predictions are added to the labeled training set. Next, the classifier is re-trained on the new enlarged labeled set and this procedure is repeated until some stopping criteria are met, such as having no unlabeled example left. The advantage of self-training algorithm is its simplicity, being at the same time one of the most efficient self-labeled algorithms. However, one disadvantage is the possible incorporation of noise examples into the labeled training set. *Co-training* is a multi-view algorithm meaning that the

feature space is divided in two conditionally independent views. Two classifiers are initially trained separately on each view using the labeled set and then following an iterative procedure each classifier adds the most confident predictions to the training set of the other. Most self-labeled methods are influenced by these two algorithms while some of them are also based on ensemble techniques. Democratic-Co learning, SETRED, Tri-training, Co-Forest and Co-Bagging are some other efficient self-labeled algorithms proposed in the literature.

Democratic-Co learning [24] is a single view algorithm utilizing multiple classifiers for predicting the values of unlabeled examples incorporating a majority voting strategy and a mean confidence measurement for the majority and the minority of classifiers which disagree. *SETRED* is a self-labeled algorithm [25] which incorporates data editing in the self-training framework in order to learn actively from the self-labeled examples. More analytically, data editing is a method which improves the quality of the training set by identifying and eliminating the mislabeled examples acquired from a self-training procedure with the help of some local information in a neighborhood graph. *Tri-training* algorithm [26] constitutes a single view algorithm which utilizes three classifiers for teaching each other based on a majority voting strategy avoiding by this way the confidence measuring of the labeling of the classifiers since this process is sometimes time-consuming and quite complicated. In particular, three classifiers are trained on data subsets generated through bootstrap sampling from the initial training set, which are then used to label the instances of the unlabeled set. If two classifiers agree on predicting a value for an unlabeled example, then this is labeled for the third too. Motivated by the previous works, Li & Zhou [27] proposed *Co-Forest* algorithm. This algorithm utilizes Random trees, which are trained on bootstrap data from the dataset assigning a few unlabeled examples to each Random tree. The final decision is made by a majority voting. Two advantages comparing with the rest self-labeled algorithms are that no physical connection between the attributes is required because of the random selection of features from the basic feature vector and the reduced ripples of its performance in the case of small number of labeled examples provided. *Co-Bagging* algorithm [28] utilizes several base classifiers trained on bootstrap data created by random resampling with replacement from the training set. Each sample contains around 2/3 of the training set. One advantage is that this algorithm works well for unstable learning algorithms where a small change in the input training data can cause a significant big change in the output hypothesis.

4. iCST-Voting

During the last decade, research focused on incorporating ensemble learning techniques in the self-labeled framework in order to build powerful classifiers. Recently, Livieris et al. [29] utilized an ensemble of classifiers as base learners to increase the efficiency of semi-supervised self-labeled methods. In another study, Livieris et al. [17] proposed the *CST-Voting* algorithm which exploits the individual predictions of the three of the most efficient self-labeled algorithms, i.e Co-training, Self-training and Tri-training utilizing a simple majority voting.

Motivated by the previous works, we attempt to go a step further by hybridizing these approaches and improving the classification efficiency of CST-Voting by incorporating an ensemble of classifiers as a base learner in all its component self-labeled algorithms. This is the main novelty of the present work over CST-Voting [17], resulting in increased prediction accuracy of iCST-Voting over CST-Voting. A high-level description of this framework is presented in Table 1 which is composed of two phases: *Training* and *Voting* phase.

In the Training phase, the self-labeled algorithms which constitute iCST-Voting are trained utilizing an ensemble of supervised classifiers E as base learner and the same labeled L and unlabeled U datasets (Steps 1-3). Subsequently, the trained classifiers $C_{\text{Co}}, C_{\text{Self}}$ and C_{Tri} are constructed using Co-training, Self-training and Tri-training algorithms, respectively.

Next, in the Voting phase, each trained self-labeled classifier is applied on each unlabeled example x of the test set T. Let y_C, y_S and y_T be the hypothesis of the classifiers C_{Co}, C_{Self} and C_{Tri} on x, respectively (Steps 6-8). The final hypothesis y^* on x is defined by combining the individual

178 predictions y_C, y_S and y_T utilizing a majority voting methodology (Step 9).
179

Table 1. iCST-Voting framework

Input: L − Set of labeled instances (Training labeled set).
U − Set of unlabeled instances (Training unlabeled set).
T − Set of unlabeled test instances (Testing set).
E − Ensemble of supervised base learners.

Output: T^* − Set of labeled instances of the testing set.

/* Phase I: Training */
[1]: $C_{\text{Co}} \leftarrow \text{Co-training}(L, U, E)$. (*Trained classifier using Co-training*)
[2]: $C_{\text{Self}} \leftarrow \text{Self-training}(L, U, E)$. (*Trained classifier using Self-training*)
[3]: $C_{\text{Tri}} \leftarrow \text{Tri-training}(L, U, E)$. (*Trained classifier using Tri-training*)

/* Phase II: Voting */
[4]: Set $T^* = \emptyset$.
[5]: **for each** $x \in T$ **do**
[6]: $y_C \leftarrow C_{\text{Co}}(x)$. (*Apply classifier C_{Co} on instance x*)
[7]: $y_S \leftarrow C_{\text{Self}}(x)$. (*Apply classifier C_{Self} on instance x*)
[8]: $y_T \leftarrow C_{\text{Tri}}(x)$. (*Apply classifier C_{Tri} on instance x*)
[9]: $y^* \leftarrow MajorityVoting\{y_C, y_S, y_T\}$.
[10]: Insert pair (x, y^*) in T^*.
[11]: **end for**

180 ## 5. Numerical experiments

181 In this section, we present a series of experiments in order to evaluate the performance of the
182 proposed algorithm iCST-Voting for gender classification from voice. The implementation codes were
183 written in Java, using the WEKA 3.9 Machine Learning Toolkit [30].
184 Our numerical experiments took place in two distinct phases: In the first phase (Section 5.2)
185 we evaluate the performance of the iCST-Voting, against its component self-labeled algorithms:
186 Self-training, Co-training and Tri-training and the state-of-the-art self-labeled algorithms: SETRED,
187 Co-Bagging, Democratic-Co learning and Co-Forest; while in the second phase (Section 5.3), we
188 compare the performance of the proposed algorithm iCST-Voting against classical supervised
189 algorithms. Table 2 reports the configuration parameters of all evaluated self-labeled algorithms.
190 In our original experiments, we have utilized several choices of parameter values, nevertheless
191 the classification performance was usually worse and only in few cases there was a negligible and
192 marginal improvement. The performance of the classification algorithms was evaluated using the
193 performance metrics F-measure (F_1) and Accuracy (Acc).
194

(continued).

Algorithm	Type	Parameters
Co-Bagging	Self-labeled - Multiple classifier	MaxIter = 40. pool U = 100. Committee members = 3. Ensemble learning = Bagging.
Democratic-Co	Self-labeled - Multiple classifier	Classifiers = kNN, C4.5, NB.
Co-Forest	Self-labeled - Multiple classifier	Number of Random Forest classifiers = 6. Threshold = 0.75.

195

Table 2. Parameter specification for all compared self-labeled algorithm used in the experimentation

Algorithm	Type	Parameters
Self-training	Self-labeled - Single classifier	MaxIter = 40. $c = 0.95$.
Co-training	Self-labeled - Multiple classifier	MaxIter = 40. Initial unlabeled pool = 75.
Tri-training	Self-labeled - Multiple classifier	No parameters specified.
SETRED	Self-labeled - Single classifier	MaxIter = 40. Threshold = 0.1.

5.1. Dataset

The efficiency of the all self-labeled algorithms was evaluated using the Voice gender dataset and the Deterding dataset [31].

- *Voice gender dataset*[1]. This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. It consists of 3168 recorded voice samples, collected from male and female speakers. The voice samples were collected from
  - The Harvard-Haskins Database of Regularly-Timed Speech.
  - Telecommunications & Signal Processing Laboratory Speech Database at McGill University.
  - VoxForge Speech Corpus.
  - Festvox CMU-ARCTIC Speech Database at Carnegie Mellon University.

  Each voice sample is stored as a .WAV file, which is then pre-processed by acoustic analysis in software `R` using the `seewave` and `tuneR` packages, with an analyzed frequency range of 0hz-280hz (human vocal range).
- *Deterding dataset*[2]. This dataset consists of the steady-state portion of 11 vowels in British English, spoken in the context of h*d. The recorded speech samples were low-pass filtered at 4.7KHz before being digitised at 10KHz with a 12-bit resolution and the steady-state portion of the vowel in each utterance was partitioned into six 512 Hamming windowed segments. Next, linear predictive analysis was performed and the linear prediction reflection coefficients were calculated and used to generate 10 log area parameters. These parameters were recorded and constituted a 10-dimensional input space. Totally, the dataset consists of 990 recorded voice samples, collected from 528 male and 462 female speakers.

All algorithms were evaluated using the stratified 10-fold cross-validation and in order to study the influence of the amount of labeled data, five different ratios (R) of the training data were used, i.e., 10%, 20%, 30%, 40% and 50%.

5.2. Performance evaluation of iCST-Voting against state-of-the-art self-labeled algorithms

Next, we focus our interest on the experimental analysis for evaluating the classification performance of iCST-Voting against the most efficient and frequently utilized self-labeled algorithms. The proposed algorithm iCST-Voting utilizes an ensemble classifier as base learner which combines the individual predictions of Sequential Minimum Optimization (SMO) algorithm [32], kNN algorithm [33] and C4.5 decision tree algorithm [34] utilizing a majority voting. These supervised

[1] https://www.kaggle.com/primaryobjects/voicegender/home
[2] https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Vowel+Recognition+-+Deterding+Data)

classifiers probably constitute the most effective and popular machine learning algorithms for classification problems [35]. Additionally, in order to present a complete performance evaluation of iCST-Voting, similar to [29], Self-training, Co-training and Tri-training utilized as base learner, the ensemble classifier used by iCST-Voting. The configuration parameters for all supervised classifiers were set as in [14] which are also reported in Table 3 for completeness. It worth noticing that SMO was implemented using Pearson VII function-based universal kernel instead of the classical polynomial kernel, which significantly improved its performance.

Table 3. Parameter specification for all the base learners used in the experimentation

Algorithm	Parameters
SMO	$C = 1.0$, Epsilon $= 1.0 \times 10^{-12}$, Tolerance parameter $= 0.001$, Kernel type = Pearson VII function-based universal kernel, Fit logistic models = true.
C4.5	Confidence level: $c = 0.25$, Mininum number of item-sets per leaf: $i = 2$, Prune after the tree building.
kNN	Number of neighbors = 3, Euclidean distance.

Table 4 presents the performance evaluation of the iCST-Voting against the state-of-the-art self-labeled algorithms Self-training, Co-training, Tri-training, SETRED, Co-Bagging, Democratic-Co learning, Co-Forest on Voice gender dataset. Notice that the highest classification performance for each labeled ratio and performance metric is highlighted in bold. It is worth mentioning that the aggregated results demonstrate that iCST-Voting is the most efficient and robust method, independent of the utilized ratio of labeled instances in the training set.

Table 4. Performance evaluation iCST-Voting against state-of-the-art self-labeled algorithms on Voice gender dataset

Algorithm	$R = 10\%$		$R = 20\%$		$R = 30\%$		$R = 40\%$		$R = 50\%$	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
Self-training	97.29%	97.29%	97.35%	97.35%	97.73%	97.73%	97.76%	97.76%	97.76%	97.76%
Co-training	97.01%	97.00%	97.04%	97.03%	97.16%	97.16%	97.19%	97.19%	97.29%	97.29%
Tri-training	97.55%	97.54%	97.64%	97.63%	97.83%	97.82%	97.86%	97.85%	97.95%	97.95%
SETRED	94.94%	94.92%	95.18%	95.17%	95.64%	95.61%	95.73%	95.71%	95.73%	95.71%
Co-Bagging	97.41%	97.41%	97.41%	97.41%	97.47%	97.47%	97.47%	97.47%	97.57%	97.57%
Democratic Co	94.86%	94.92%	95.69%	95.74%	96.22%	96.24%	96.59%	96.62%	97.23%	97.19%
Co-Forest	97.17%	97.16%	97.20%	97.19%	97.30%	97.29%	97.36%	97.35%	97.48%	97.47%
iCST-Voting	**97.92%**	**97.92%**	**97.98%**	**97.98%**	**98.14%**	**98.14%**	**98.24%**	**98.23%**	**98.43%**	**98.42%**

Table 5 reports the performance of Self-training, Co-training, Tri-training, SETRED, Co-Bagging, Democratic-Co learning, Co-Forest and iCST-Voting on Deterding dataset. As above mentioned, the accuracy measure of the best performing algorithm is highlighted in bold. Similar observations can be made with the previous benchmark. Clearly, iCST-Voting was by far the most efficient and robust method, demonstrating 1.21%-5.66% better classification accuracy, independent of the utilized ratio of labeled instances in the training set.

Table 5. Performance evaluation iCST-Voting against state-of-the-art self-labeled algorithms on Deterding dataset

Algorithm	$R = 10\%$		$R = 20\%$		$R = 30\%$		$R = 40\%$		$R = 50\%$	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
Self-training	91.50%	90.81%	91.89%	91.21%	92.18%	91.52%	92.16%	91.52%	92.38%	91.72%
Co-training	90.57%	89.90%	90.65%	90.00%	91.06%	90.40%	89.93%	89.09%	90.98%	90.30%
Tri-training	91.43%	90.71%	91.79%	91.11%	91.99%	91.31%	92.08%	91.41%	92.26%	91.62%
SETRED	91.14%	90.30%	91.65%	90.91%	91.65%	90.91%	92.18%	91.52%	92.72%	92.12%
Co-Bagging	89.47%	88.69%	90.16%	89.39%	90.16%	89.39%	90.70%	90.00%	91.10%	90.40%
Democratic-Co	90.45%	89.39%	90.45%	89.39%	91.04%	90.10%	91.76%	90.91%	92.10%	91.31%
Co-Forest	88.47%	87.58%	88.83%	87.98%	88.93%	88.08%	89.01%	88.18%	89.10%	88.28%
iCST-Voting	**93.02%**	**92.42%**	**93.32%**	**92.73%**	**93.32%**	**92.73%**	**93.61%**	**93.03%**	**94.45%**	**93.94%**

The statistical comparison of several classification algorithms over multiple datasets is fundamental in the area of machine learning and it is frequently performed by means of a statistical test [17,18]. Since our motivation stems from the fact that we are interested in evaluating the rejection of the hypothesis that all the algorithms perform equally well for a given level based on their classification accuracy and highlighting the existence of significant differences between our proposed algorithm and the classical self-labeled algorithms, we utilized the non-parametric Friedman Aligned Ranking (FAR) [36] test. Additionally, in order to identify which algorithms report significant differences, the Finner test [37] is applied as a post-hoc procedure. It is worth mentioning that the control algorithm for the post-hoc test is determined by the best (lowest) ranking obtained in each FAR test. Furthermore, the adjusted p-value with Finner's test (p_F) was presented based on the corresponding control algorithm at the $\alpha = 0.05$ level of significance while the post-hoc test test rejects the hypothesis of equality when the value of p_F is less than the value of a.

Table 6 presents the information of the statistical analysis performed by nonparametric multiple comparison procedures for Self-training, Co-training, Tri-Training, CST-Voting and iCST-Voting. The interpretation of Table 6 illustrates that the proposed iCST-Voting algorithm reports the highest probability-based ranking by statistically presenting better results, outperforming the rest self-labeled algorithms.

Table 6. Friedman Aligned Ranking (FAR) test and Finner post-hoc test for Self-training, Co-training, Tri-training, SETRED, Co-Bagging, Democratic-Co learning, Co-Forest and iCST-Voting

Algorithm	FAR	Finner Post-Hoc Test	
		p_F-value	Null Hypothesis
iCST-Voting	5.7		-
Tri-training	22.8	0.048123	rejected
Self-training	23.85	0.043013	rejected
Co-training	53.7	0.035975	rejected
SETREG	48.6	0.028885	rejected
Co-Bagging	50.05	0.021743	rejected
Democratic-Co	58.6	0.014548	rejected
Co-Forest	60.7	0.007301	rejected

5.3. Performance evaluation of iCST-Voting against classical supervised algorithms

In the sequel, we evaluate the classification performance of iCST-Voting against the classical supervised classification algorithms: SMO, kNN and C4.5. Moreover, we compare the performance of iCST-Voting against the ensemble of classifiers (Voting) which combines the individual predictions of the supervised classifiers utilizing a majority voting strategy.

Tables 7 and 8 report the performance of the supervised algorithms SMO, *k*NN, C4.5, Voting against iCST-Voting on Voice gender dataset and Deterding dataset, respectively, trained with different amounts of labeled data. As above mentioned, the highest classification performance for each labeled ratio and performance metric is highlighted in bold. The aggregated results show that iCST-Voting was by far the most efficient algorithm since it illustrates the highest classification performance, independent of the utilized ratio of labeled instances in the training set. More specifically, iCST-Voting outperforms all classical supervised algorithms, regarding both performance metrics and datasets.

Table 7. Performance evaluation iCST-Voting against state-of-the-art supervised algorithms on Voice gender dataset, relative to each labeled ratio

Algorithm	$R = 10\%$		$R = 20\%$		$R = 30\%$		$R = 40\%$		$R = 50\%$	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
SMO	96.00%	95.92%	96.84%	96.84%	97.40%	97.37%	97.96%	97.95%	97.98%	97.98%
*k*NN	95.77%	95.57%	96.25%	96.22%	97.08%	97.05%	97.95%	97.95%	97.55%	97.54%
C4.5	94.79%	94.97%	95.64%	95.57%	94.32%	94.32%	95.77%	95.74%	96.19%	96.21%
Voting	96.67%	96.53%	96.55%	96.53%	97.18%	97.16%	97.89%	97.87%	97.61%	97.60%
iCST-Voting	**97.92%**	**97.92%**	**97.98%**	**97.98%**	**98.14%**	**98.14%**	**98.24%**	**98.23%**	**98.43%**	**98.42%**

Table 8. Performance evaluation iCST-Voting against state-of-the-art supervised algorithms on Deterding dataset, relative to each labeled ratio

Algorithm	$R = 10\%$		$R = 20\%$		$R = 30\%$		$R = 40\%$		$R = 50\%$	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
SMO	92.05%	90.89%	92.74%	92.42%	93.16%	92.64%	93.42%	93.01%	93.80%	93.35%
*k*NN	88.12%	87.78%	92.62%	91.95%	91.77%	90.93%	92.60%	92.18%	93.99%	93.76%
C4.5	79.56%	76.78%	82.37%	80.66%	87.47%	86.49%	87.31%	86.90%	92.24%	91.73%
Voting	86.21%	84.78%	90.33%	89.32%	91.73%	90.92%	92.02%	92.69%	93.92%	93.56%
iCST-Voting	**93.02%**	**92.42%**	**93.32%**	**92.73%**	**93.32%**	**92.73%**	**93.61%**	**93.03%**	**94.45%**	**93.94%**

Table 9 presents the performance of the supervised algorithms SMO, *k*NN, C4.5, Voting and iCST-Voting, regarding both classification benchmarks. It is worth noticing that all supervised classifiers were trained using 100% of the training data while iCST-Voting utilized only 50% of the training data. Relative to Voice gender dataset, the proposed algorithm iCST-Voting presents the best performance, outperforming all supervised algorithms. Regarding the Deterding dataset, the best performance is achieved by *k*NN while the proposed algorithm iCST-Voting slightly outperforms SMO and C4.5.

Table 9. Performance evaluation iCST-Voting against state-of-the-art supervised algorithms trained with 100% of training set on Voice gender and Deterding datasets

Algorithm	Voice gender dataset		Deterding dataset	
	F_1	Acc	F_1	Acc
SMO	98.38%	98.39%	94.45%	93.93%
*k*NN	98.12%	98.11%	**99.44%**	**99.39%**
C4.5	96.83%	96.84%	94.05%	93.64%
Voting	97.94%	97.95%	97.38%	97.17%
iCST-Voting	**98.43%**	**98.42%**	94.45%	93.94%

6. Conclusions

In this work, we utilized a semi-supervised algorithm, called iCST-Voting for the gender recognition by voice. The proposed algorithm constitutes an ensemble of the most popular self-labeled algorithms i.e. Self-training, Co-training and Tri-training utilizing as base learner an ensemble of classifiers. The contribution of our approach as compared to other related approaches has to do with the fact that we utilize an ensemble of classifiers as base learner instead of single learners normally used in self-labeled algorithms. Our preliminary numerical results and the presented statistical analysis demonstrate the efficiency of the proposed algorithm for the gender recognition by voice compared against state-of-the-art self-labeled. Moreover, it presents competitive and sometimes better classification performance than classical supervised algorithms. Therefore, we conclude that reliable, stable and robust prediction models could be developed by the adaptation of ensemble techniques in the semi-supervised learning framework.

In order to resolve possible scalability issues we run a set of experiments using two large datasets: the Pneumonia dataset [38] and CT Medical dataset [39] and found that in our proposed approach there was no degradation in accuracy of prediction. Only an expected small increase in training time was noticed. Therefore, we can safely say that our proposed approach performs equally well even when input data increases drastically. Moreover, it is worth noticing that we understand the limitations imposed on the generalizability of the presented results due to the use of the only two freely available data. We certainly intend to investigate this further in the near future.

Our future work is focused on improving the prediction accuracy of our framework by combining self-labeled algorithms with more sophisticated and theoretically motivated ensemble learning methodologies. Additionally, another interesting aspect is concentrating on extending our framework for handling big data with traditional technique and platforms such as [40,41]. Finally, since our numerical experiments are quite promising, we intend to focus on expanding our experiments and applying further the proposed algorithm to several audio datasets for speaker recognition.

Author Contributions: I.E. Livieris, E. Pintelas and P. Pintelas conceived of the idea, designed and performed the experiments, analyzed the results, drafted the initial manuscript and revised the final manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

1. Pahwa, A.; Aggarwal, G. Speech feature extraction for gender recognition. *International Journal of Image, Graphics and Signal Processing* **2016**, *8*, 17.
2. Gamit, M.R.; Dhameliya, K.; Bhatt, N.S. Classification Techniques for Speech Recognition: A Review. *International Journal of Emerging Technology and Advanced Engineering* **2015**, *5*, 58–63.
3. Yasmin, G.; Dutta, S.; Ghosal, A. Discrimination of male and female voice using occurrence pattern of spectral flux. Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017 International Conference on. IEEE, 2017, pp. 576–581.
4. Bisio, I.; Lavagetto, F.; Marchese, M.; Sciarrone, A.; Frà, C.; Valla, M. Spectra: A speech processing platform as smartphone application. 2015 IEEE International Conference on Communications. IEEE, 2015, pp. 7030–7035.
5. Wang, W.C.; Pestana, M.H.; Moutinho, L. The Effect of Emotions on Brand Recall by Gender Using Voice Emotion Response with Optimal Data Analysis. In *Innovative Research Methodologies in Management*; Springer, 2018; pp. 103–133.
6. Holzinger, A. Introduction to machine learning and knowledge extraction (MAKE). *Machine Learning and Knowledge Extraction* **2017**, *1*, 1–20.
7. Ferri, M. Why topology for machine learning and knowledge extraction? *Machine Learning and Knowledge Extraction* **2018**, *1*, 115–120.

8. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann, 2016.
9. Buyukyilmaz, M.; Cibikdiken, A.O. Voice Gender Recognition Using Deep Learning. *Advances in Computer Science Research* **2016**, *58*, 409–411.
10. Maka, T.; Dziurzanski, P. An analysis of the influence of acoustical adverse conditions on speaker gender identification. Pacific Voice Conference (PVC), 2014 XXII Annual. IEEE, 2014, pp. 1–4.
11. Zvarevashe, K.; Olugbara, O.O. Gender Voice Recognition Using Random Forest Recursive Feature Elimination with Gradient Boosting Machines. 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems. IEEE, 2018, pp. 1–6.
12. Harb, H.; Chen, L. A general audio classifier based on human perception motivated model. *Multimedia Tools and Applications* **2007**, *34*, 375–395.
13. Vogt, T.; André, E. Improving automatic emotion recognition from speech via gender differentiation. Proceddings Language Resources and Evaluation Conference, 2006.
14. Triguero, I.; García, S.; Herrera, F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems* **2015**, *42*, 245–284.
15. Silva, N.F.F.D.; Coletta, L.F.; Hruschka, E.R. A survey and comparative study of tweet sentiment analysis via semi-supervised learning. *ACM Computing Surveys (CSUR)* **2016**, *49*, 15.
16. Hajighorbani, M.; Hashemi, S.R.; Broumandnia, A.; Faridpour, M. A review of some semi-supervised learning methods. IEEE-2016, First International Conference on New Research Achievements in Electrical and Computer Engineering, 2016.
17. Livieris, I.E.; Kanavos, A.; Tampakas, V.; Pintelas, P. An ensemble SSL algorithm for efficient chest x-ray image classification. *Journal of Imaging* **2018**, *4*.
18. Livieris, I.E.; Kiriakidou, N.; Kanavos, A.; Tampakas, V.; Pintelas, P. On ensemble SSL algorithms for credit scoring problem. *Informatics* **2018**, *5*.
19. Přibil, J.; Přibilová, A.; Matoušek, J. GMM-based speaker gender and age classification after voice conversion. Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016 First International Workshop on. IEEE, 2016, pp. 1–5.
20. Přibil, J.; Přibilová, A.; Matoušek, J. GMM-based speaker age and gender classification in Czech and Slovak. *Journal of Electrical Engineering* **2017**, *68*, 3–12.
21. Přibil, J.; Přibilová, A.; Matoušek, J. Experiment with GMM-based artefact localization in Czech synthetic speech. International Conference on Text, Speech, and Dialogue. Springer, 2015, pp. 23–31.
22. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd annual meeting of the association for computational linguistics, 1995, pp. 189–196.
23. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.
24. Zhou, Y.; Goldman, S. Democratic co-learning. 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2004, pp. 594–602.
25. Li, M.; Zhou, Z. SETRED: Self-training with editing. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2005, pp. 611–621.
26. Zhou, Z.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering* **2005**, *17*, 1529–1541.
27. Li, M.; Zhou, Z. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **2007**, *37*, 1088–1098.
28. Hady, M.F.A.; Schwenker, F. Combining committee-based semi-supervised learning and active learning. *Journal of Computer Science and Technology* **2010**, *25*, 681–698.
29. Livieris, I.E.; Drakopoulou, K.; Mikropoulos, T.A.; Tampakas, V.; Pintelas, P. An ensemble-based semi-supervised approach for predicting students' performance. In *Research on e-Learning and ICT in Education*; Springer, 2018; pp. 25–42.
30. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA data mining software: An update. *SIGKDD explorations newsletters* **2009**, *11*, 10–18.
31. Deterding, D.H. Speaker normalization for automatic speech recognition. PhD thesis, University of Cambridge, 1989.

32. Platt, J. *Advances in Kernel Methods - Support Vector Learning*; MIT Press: Cambridge, Massachusetts, 1998.
33. Aha, D. *Lazy learning*; Dordrecht: Kluwer academic publishers, 1997.
34. Quinlan, J. *C4.5: Programs for machine learning*; Morgan Kaufmann: San Francisco, 1993.
35. Wu, X.; Kumar, V.; Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.; Ng, A.; Liu, B.; Yu, P.; Zhou, Z.; Steinbach, M.; Hand, D.; Steinberg, D. Top 10 algorithms in data mining. *Knowledge and information systems* **2008**, *14*, 1–37.
36. Hodges, J.; Lehmann, E. Rank methods for combination of independent experiments in analysis of variance. *The annals of mathematical statistics* **1962**, *33*, 482–497.
37. Finner, H. On a monotonicity problem in step-down multiple test procedures. *Journal of the American statistical association* **1993**, *88*, 920–923.
38. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; others. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131.
39. Albertina, B.; Watson, M.; Holback, C.; Jarosz, R.; Kirk, S.; Lee, Y.; Lemmerman, J. Radiology data from the cancer genome atlas lung adenocarcinoma [tcga-luad] collection. *The Cancer Imaging Archive* **2016**.
40. Anagnostopoulos, I.; Zeadally, S.; Exposito, E. Handling big data: research challenges and future directions. *The Journal of Supercomputing* **2016**, *72*, 1494–1516.
41. Kolias, V.; Kolias, C.; Anagnostopoulos, I.; Kayafas, E. RuleMR: Classification rule discovery with MapReduce. 2014 IEEE International Conference on Big Data. IEEE, 2014, pp. 20–28.