UDC 519.87

Vestnik SibGAU
Vol. 16, No. 1, P. 28–34

ROBUST AND RELIABLE TECHNIQUES FOR SPEECH-BASED EMOTION RECOGNITION

C. Yu. Brester[1*], O. E. Semenkina[1], M. Yu. Sidorov[2]

[1]Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
[2]Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany
[*]E-mail: christina.bre@yandex.ru

One of the crucial challenges related to the spacecraft control is the monitoring of the mental state of crew members as well as operators of the flight control centre. In most cases, visual information is not sufficient, because spacemen are trained to cope with feelings and not to express emotions explicitly. In order to identify the genuine mental state of a crew member, it is reasonable to engage the acoustic characteristics obtained from speech signals presenting voice commands during the spacecraft control and interpersonal communication. Human emotion recognition implies flexible algorithmic techniques satisfying the requirements of reliability and fast operation in real time. In this paper we consider the heuristic feature selection procedure based on the self-adaptive multi-objective genetic algorithm that allows the number of acoustic characteristics involved in the recognition process to be reduced. The effectiveness of this approach and its robustness property are revealed in experiments with various classification models. The usage of this procedure leads to a reduction of the feature space dimension by a factor of two (from 384 to approximately 180 attributes), which means decreasing the time resources spent by the recognition algorithm. Moreover, it is proposed to implement some algorithmic schemes based on collective decision making by the set of classifiers (Multilayer Perceptron, Support Vector Machine, Linear Logistic Regression) that permits the improvement of the recognition quality (by up to 10% relative improvement). The developed algorithmic schemes provide a guaranteed level of effectiveness and might be used as a reliable alternative to the random choice of a classification model. Due to the robustness property the heuristic feature selection procedure is successfully applied on the data pre-processing stage, and then the approaches realizing the collective decision making schemes are used.

Keywords: emotion recognition, adaptive multi-objective genetic algorithm, classifier, collective decision making.

Вестник СибГАУ
Т. 16, № 1. С. 28–34

РОБАСТНЫЕ И НАДЕЖНЫЕ ПОДХОДЫ К РАСПОЗНАВАНИЮ ЭМОЦИЙ ПО РЕЧИ

К. Ю. Брестер[1*], О. Э. Семенкина[1], М. Ю. Сидоров[2]

[1]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
[2]Ульмский университет
Германия, 89081, г. Ульм, аллея им. Альберта Эйнштейна, 43
[*]E-mail: christina.bre@yandex.ru

Контроль психоэмоционального состояния членов экипажа космического корабля, а также операторов центра управления полетами является одной из ключевых задач, требующих решения в онлайн-режиме. Нередко визуальный контроль может быть недостаточным, поскольку экипаж обучен владеть собой и не выражать эмоций в явном виде. Для определения более точного психологического портрета возможно использование акустических характеристик речевых сигналов, фиксируемых в ходе управления космическим аппаратом (голосовыми командами) и обычной коммуникации. Распознавание эмоций человека в ходе коммуникации с интеллектуальными диалоговыми системами предполагает наличие гибкого алгоритмического аппарата, отвечающего требованиям надежности и быстродействия при функционировании в режиме реального времени. Рассматривается эвристическая процедура извлечения информативных признаков, позволяющая существенно сократить число акустических характеристик, используемых алгоритмами распознавания. Эффективность данного подхода исследуется в сочетании с различными классификационными моделями, благодаря чему демонстрируется свойство робастности. Применение указанной процедуры позволяет снизить размерность

признакового пространства в два раза (с 384 до приблизительно 180 атрибутов), что сопряжено с сокращением временных ресурсов, затрачиваемых алгоритмом распознавания. Кроме того, было предложено несколько алгоритмических схем, основанных на коллективном принятии решений набором классификаторов, что позволило существенно повысить качество распознавания (приблизительно до 10 % для одной из баз данных). Разработанные алгоритмические схемы обеспечивают гарантированный уровень эффективности и являются надежной альтернативой произвольному выбору классификационной модели. Благодаря свойству робастности, эвристическая процедура отбора информативных признаков была успешно использована на этапе предобработки данных с последующим применением подходов, реализующих механизмы коллективного принятия решений.

Ключевые слова: распознавание эмоций, адаптивный многокритериальный генетический алгоритм, классификатор, коллективное принятие решений.

Introduction. During monitoring of the spacecraft flight, it is essential to assess the astronaut abilities to provide the reliable control with sober mind. In most cases, instability of the emotions state may prevent the crew from making a right decision. Moreover, the usage of visual information is likely to be less relevant for this purpose because astronauts are trained to hide their genuine emotions and keep calm explicitly. Therefore, it is reasonable to recognize their emotional state based on speech signals, in particular, based on voice commands while operating with the spacecraft and interpersonal communication.

Although many good results have already been achieved in the sphere of emotion recognition, there are a number of open questions. Some of them are about the development of effective classification methods that should be applied to this problem [1]. Others pertain to extracting acoustic characteristics from speech signals [2; 3] or selecting the set of relevant features from databases [4].

At the "INTERSPEECH 2009 Emotion Challenge" an appropriate set of acoustic characteristics used to describe any speech signal was introduced. This set of features, comprising attributes such as pitch, intensity and formants, is high-dimensional: the number of characteristics is 384. For most classifiers it is extremely difficult to make a decision based on all this input data: features might have a low variation level, correlate with each other or be measured with mistakes.

Background. During experiments it was revealed that the usage of the standard feature selection procedures (such as Principal Component Analysis (PCA)) led to the classification accuracy decreasing [5]. Therefore, to oppose this, some heuristic techniques based on the multiobjective genetic algorithm were developed.

Two main schemes of dimension reduction are realized normally to determine the relevant feature set [6]. According to the first one, it is compulsory to evaluate the effectiveness of the selected attributes with any classification model (the *wrapper* approach). The second method requires some metrics (Attribute Class Correlation, Inter- and Intra- Class Distances, Laplasian Score, Representation Entropy and the Inconsistent Example Pair measure) to be estimated and it ignores the classifier performance entirely (the *filter* approach) [7]. The details of the presented schemes and criteria introduced are described in [8].

As the feature search procedure we implement the Strength Pareto Evolutionary Algorithm (SPEA) [9] based on the Pareto dominance idea (fig. 1). It operates with a set of binary strings coding the informative features in the following way: *unit* corresponds to the relevant attribute whereas *zero* denotes the irrelevant one.

SPEA uses the outer set to preserve non-dominated solutions and genetic operators to produce new candidate solutions. Furthermore, to avoid choosing the algorithm settings we suggest applying the self-adaptive modification of SPEA. Originally, tournament selection is used; therefore only crossover and mutation should be adjusted.

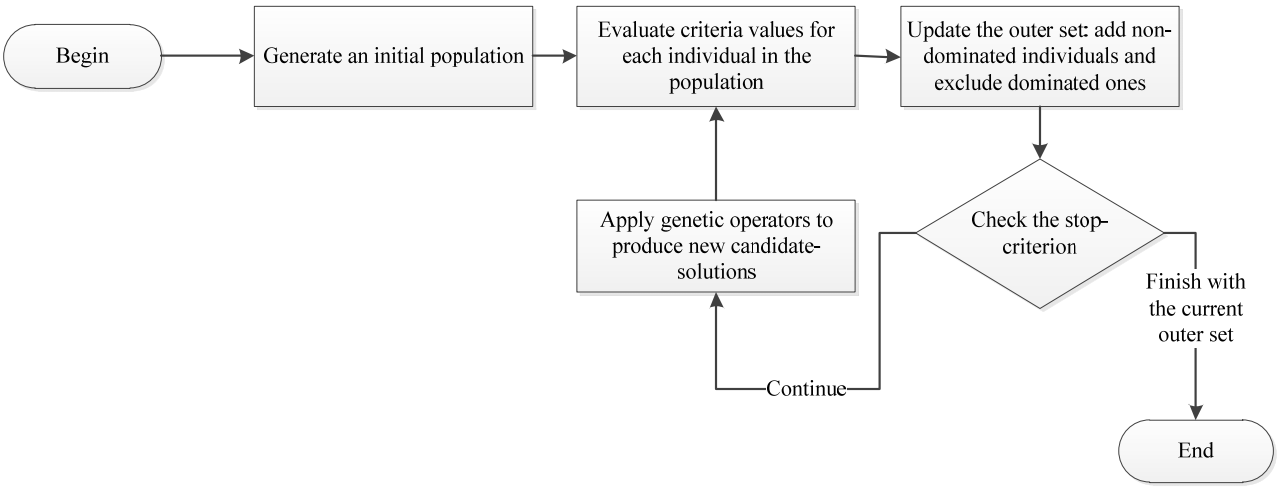

Fig. 1. The general scheme of the SPEA algorithm

We realize the self-configurable recombination operator based on the co-evolution idea [10]: the population is divided into parts and each part is derived with a particular type of crossover that is conventionally one-point, two-point or uniform. We also estimate the fitness value for each variant of recombination that implies the number of non-dominated individuals generated with it. Based on these rates, resources are reallocated in every *T*-generation. The self-adaptive mutation operator is based on the idea proposed by Daridi [11]. At every generation mutation probability is recalculated according to the heuristic rule.

The research conducted [8] has exposed that the filter approach permits the achievement of better results in the sense of classification accuracy, whereas the usage of the wrapper approach decreases the number of features significantly.

Due to the independency of the filter approach from classification models it might be supposed that this feature selection procedure should be rather effective in combination with various classifiers. Therefore in this paper we explore the *robustness property* of the filter technique. In other words, it is necessary to consider a number of classification models and check whether this method is effective for most of them or not.

Moreover, it is hardly ever possible for the online dialogue systems to vary classifiers and determine the most effective one while interacting with a user. Consequently, some general approaches based on involving different models should be elaborated. In this research we propose three schemes of taking into account predictions of different classifiers and producing the collective decision. The effectiveness of these algorithmic schemes is investigated on both baseline and reduced feature sets of emotion recognition problems.

Databases description. In the study a number of speech databases have been used and this section provides their brief description.

The *Berlin* emotional database [12] was recorded at the Technical University of Berlin and consists of labelled emotional German utterances which were spoken by 10 actors (5 female). Each utterance has one of the following emotional labels: neutral, anger, fear, joy, sadness, boredom or disgust.

The *VAM* database [13] was created at Karlsruhe University and consists of utterances extracted from the popular German talk-show “Vera am Mittag” (Vera in the afternoon). The emotional labels of the first part of the corpus (speakers 1–19) were given by 17 human evaluators and the rest of the utterances (speakers 20–47) were labelled by 6 annotators on a 3-dimensional emotional basis (valence, activation and dominance). To produce the labels for the classification task we have used just a valence (or evaluation) and an arousal axis. The corresponding quadrant (anticlockwise, starting in the positive quadrant, and assuming arousal as abscissa) can also be assigned emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene.

The *UUDB* (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database [14] consists of spontaneous Japanese human-human speech. The task-oriented dialogue produced by seven pairs of speakers (12 female) resulted in 4.737 utterances in total. Emotional labels for each utterance were created by three annotators on a five-dimensional emotional basis (interest, credibility, dominance, arousal, and pleasantness). For this work, only pleasantness (or evaluation) and the arousal axes are used.

There is a statistical description of the corpora used in tab. 1.

Robustness of the filter approach. The robustness property of the filter approach was investigated using a set of conventional classification models: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Linear Logistic Regression (Logit), Radial Basis Function network (RBF), Naive Bayes, Decision trees (J48), Random Forest, Bagging, Additive Logistic Regression (LogitBoost) and One Rule (OneR) [15].

For each classifier the *F-score* [16] was evaluated, first, on the baseline data set and, secondly, on the set of features selected by SPEA. Also the relative F-score improvement and the average number of extracted attributes were estimated. We implemented a 6-fold cross-validation procedure. For all of the corpora SPEA was provided with an equal number of resources (for each run 10100 candidate solutions were examined in the search space).

The results obtained are presented in tab. 2. The average number of selected features is equal to: Berlin – 182.2, VAM – 178.7, UUDB – 179.2.

Based on the experimental results we may conclude that there is no classification model which provides a lower F-score value for all of the corpora after the feature selection procedure. Moreover, for example, the *Decision Trees* model (J48) demonstrates improvement of the F-score in all of the experiments. Obviously, in some cases the dimension reduction is achieved at the detriment of the classifier performance.

Besides, there is no particular model that is equally effective for all of the databases. We may notice that F-score values vary significantly for different classifiers. Even the best model for a certain database might be the worst for another one. For instance, MLP demonstrates the highest performance on the Berlin corpus, whereas for the UUDB database it achieves the worst results (and vice versa, for the One Rule classifier). Therefore it might be reasonable to involve a number of classifiers in the decision making process in order to increase the reliability of the classification technique. Otherwise, the random choice of the classifier may lead to significant performance deterioration.

Table 1

Statistical description of the corpora used

Database	Language	Full length, min.	Number of emotions	File level duration		Notes
				Mean, sec.	Std., sec.	
Berlin	German	24.7	7	2.7	1.02	Acted
VAM	German	47.8	4	3.02	2.1	Non-acted
UUDB	Japanese	113.4	4	1.4	1.7	Non-acted

The next section contains a description of the algorithmic schemes that compose the final prediction based on decisions of a number of classifiers.

Collective decision making in the emotion recognition procedure. In this research we investigate the efficiency of three techniques that allow the predictions of different classifiers to be taken into account while making the final decision [17].

Scheme 1. For each test example it is necessary to determine k-nearest neighbours from the training data set. The prediction of the model that classifies these k-nearest neighbours correctly is used as the final decision. (If several models demonstrate equal effectiveness, choose one of them randomly).

Scheme 2. For each test example the engaged models vote for different classes according to their own predictions. The final decision is defined as a collective choice based on the majority rule.

Scheme 3. The previous scheme has one disadvantage: if the number of classes is greater than or equal to the number of classifiers involved (or the number of classifiers is even), a situation whereby several classes get the majority of votes often occurs. Therefore we combine Schemes 1 and 2 in the following way (fig. 2):

– fulfil the voting procedure as it is described in Scheme 2;

– if several classes have the maximum number of votes, apply Scheme 1.

In all these schemes there is no limitation to the number of classifiers.

Table 2

Experimental results for conventional classifiers

	Berlin			VAM			UUDB		
	F-score, %		Gain, %	F-score, %		Gain, %	F-score, %		Gain, %
	Baseline	SPEA		Baseline	SPEA		Baseline	SPEA	
MLP	**82.87**	**82.26**	–0.74	41.08	**43.05**	4.80	25.48	34.58	35.71
SVM	81.71	82.14	0.53	43.57	36.92	–15.26	35.59	33.04	–7.15
Logit	80.04	82.15	2.64	36.88	37.88	2.71	36.72	36.33	–1.06
RBF	68.93	71.59	3.85	37.87	34.47	–8.97	26.75	23.60	–11.77
Naive Bayes	66.91	67.45	0.81	40.86	42.33	3.60	36.52	36.45	–0.20
J48	50.15	51.96	3.60	36.20	37.70	4.17	38.70	**42.64**	10.18
Random Forest	54.69	73.43	34.27	**45.66**	37.08	–18.79	40.11	36.56	–8.84
Bagging	60.60	63.29	4.43	37.24	36.67	–1.54	40.94	37.08	–9.42
Logit Boost	66.66	71.21	6.82	40.06	36.14	–9.80	41.28	37.17	–9.96
OneR	29.20	29.20	0.00	33.34	33.34	0.00	**41.92**	41.56	–0.85

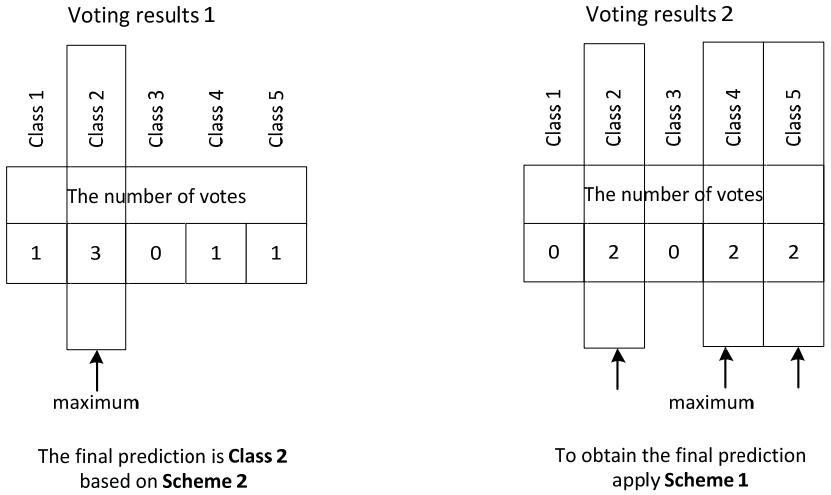

Fig. 2. An example how Scheme 3 works

Table 3

Experimental results for collective decision making schemes

	Baseline feature set			Reduced feature set		
	Scheme 1	Scheme 2	Scheme 3	Scheme 1	Scheme 2	Scheme 3
Berlin	81.18	84.01	**84.23**	79.91	**82.54**	**82.54**
VAM	42.29	**50.19**	43.69	37.99	**39.18**	**39.18**
UUDB	37.96	36.41	**39.78**	**40.43**	34.99	35.19

Analysis of the results presented in tab. 2 showed that for the used corpora Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Linear Logistic Regression (Logit) demonstrated rather high performance. Therefore it was decided to involve these classifiers in the proposed schemes of collective decision making.

We investigated the effectiveness of the developed schemes on the baseline feature set and on the set of attributes selected by SPEA (due to the robustness property of the feature selection procedure, it is also reasonable to apply these schemes including a number of models to the reduced feature set). Tab. 3 contains the F-score values obtained during the 6-fold cross-validation procedure for all of the corpora.

Based on the experimental results it might be concluded that on the set of presented databases Scheme 3 is effective for the collective classification process on the full data set as well as after the feature selection procedure.

On the Berlin database (fig. 3) all schemes demonstrate high performance. The F-score values obtained with the usage of Schemes 2 and 3 even outperform the best results achieved by MLP on the full and reduced feature sets.

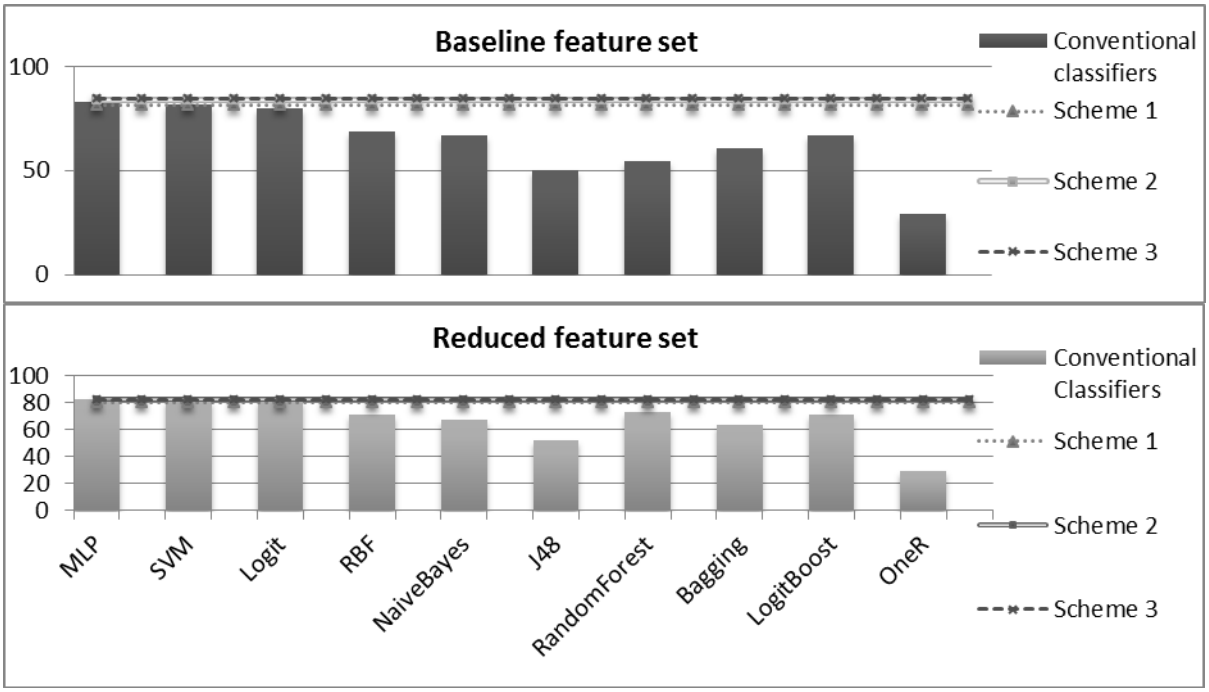

Fig. 3. Classification results for *Berlin*

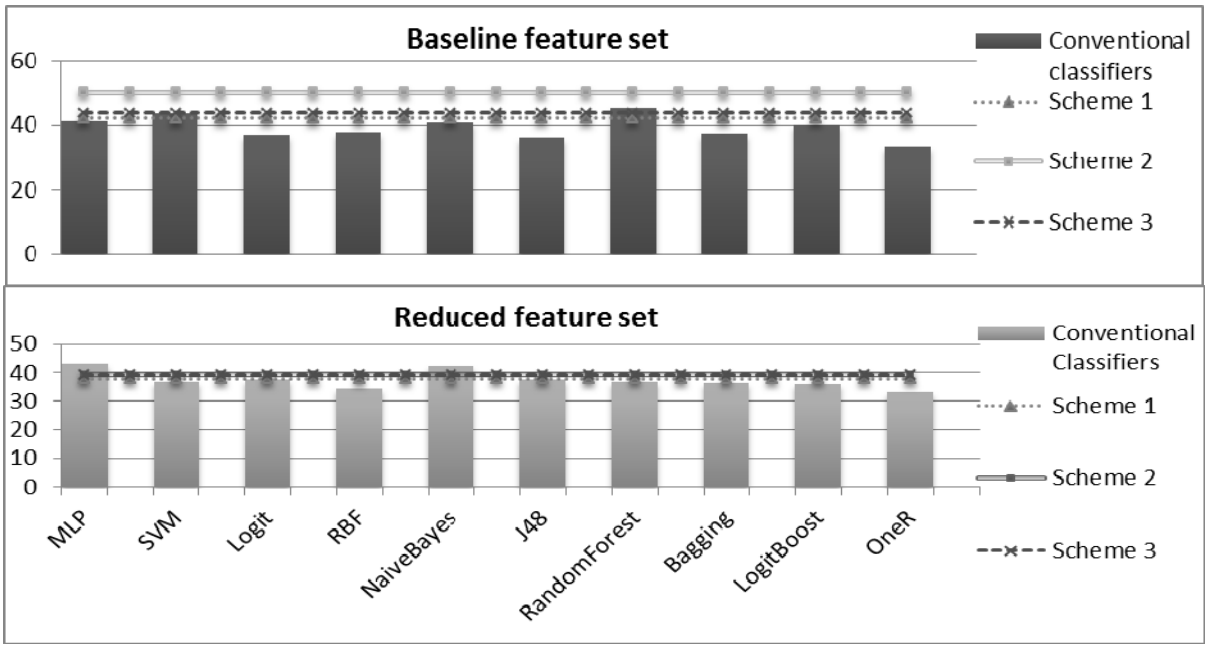

Fig. 4. Classification results for *VAM*

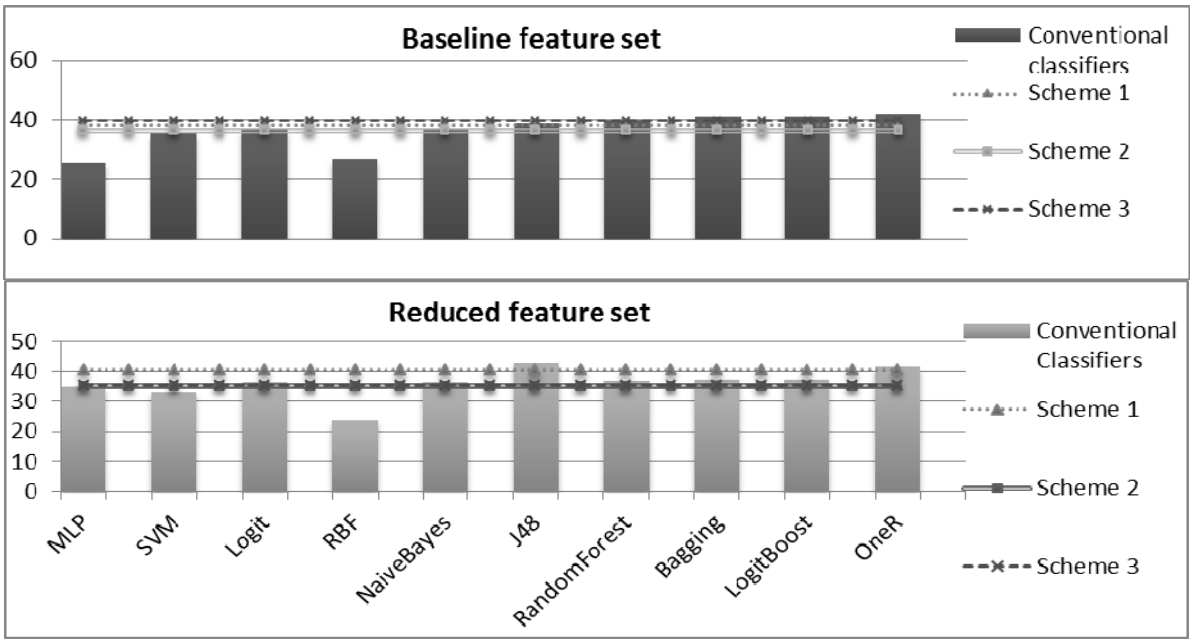

Fig. 5. Classification results for *UUDB*

In most cases the F-score values achieved by any collective decision making scheme are comparable with the best results provided by the most effective models or, at least, higher than the average F-score value obtained by conventional classifiers.

The best classification results on the VAM corpus (fig. 4) provided by Random Forest on the baseline feature set were exceeded by the application of Scheme 2 (9.93 % relative improvement). It is essential to take into account that in this case the most effective classification model (Random Forest) is not involved in the set of classifiers used in the framework of Scheme 2 (MLP, SVM, Linear Logistic Regression). Nevertheless, we attained a significantly better result with classifiers that demonstrated average effectiveness on this corpus.

Even on the UUDB corpus we obtained rather high F-score values (fig. 5), although MLP, SVM and Linear Logistic Regression demonstrated the worst results separately.

Conclusions. In this paper some effective approaches to the emotion recognition problem based on heuristic feature selection and collective decision making are considered. Due to the usage of these techniques it became possible to improve the classification results for most of the corpora (in some cases even by up to 10 % relative improvement) and, moreover, to reduce the number of features involved in the classification procedure significantly (from 384 to approximately 180).

The conducted experiments also exposed that the proposed schemes of collective choice might be effectively applied to the full data set as well as to the reduced one (after feature selection).

Although we managed to achieve some good results, there are a number of questions. The first one is related to the feature selection technique, in particular, to the introduced criteria: *whether it is reasonable to take into consideration other criteria (Laplasian Score, Representation Entropy and the Inconsistent Example Pair measure) or not? Should we engage the information about the classifier performance into the heuristic search on the stage of feature selection or ignore it totally to maintain the robustness of this approach?*

Other questions pertain to the classification models involved in the collective decision making process: *how many classifiers should we use to provide the most reliable scheme? What kind of models should it be compulsory to include in the ensemble of classifiers?*

These crucial questions will have to be elaborated in the next paper.

Acknowledgment. The research is performed with the financial support of the Ministry of Education and Science of the Russian Federation within the federal R&D program (project RFMEFI57414X0037).

Благодарности. Исследование выполняется при финансовой поддержке Министерства образования и науки Российской Федерации в рамках ФЦП ИР (проект RFMEFI57414X0037).

References

1. Sidorov M., Ultes S., Schmitt A. Emotions are a personal thing: Towards speaker-adaptive emotion recognition. *ICASSP*, 2014, p. 4803–4807.

2. Eyben F., Wöllmer M., and Schuller B. Opensmile: the munich versatile and fast opensource audio feature extractor. *Proceedings of the international conference on Multimedia*, 2010. ACM, p. 1459–1462.

3. Boersma P. Praat, a system for doing phonetics by computer. *Glot international,* 2002, no. 5(9/10), p. 341–345.

4. Sidorov M., Brester C., Minker W., Semenkin E. Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm. *LREC*, 2014, p. 3481–3485.

5. Brester C., Semenkin E., Sidorov M., Minker W. Self-adaptive multi-objective genetic algorithms for feature selection. *Proceedings of International Conference on Engineering and Applied Sciences Optimization (OPT-i'14)*, 2014, p. 1838–1846.

6. Kohavi R., John G. H. Wrappers for feature subset selection. *Artificial Intelligence,* 97, 1997, p. 273–324.

7. Venkatadri M., Srinivasa Rao K. A multiobjective genetic algorithm for feature selection in data mining. I*nternational Journal of Computer Science and Information Technologies*, 2010, vol. 1, no. 5, p. 443–448.

8. Brester C., Sidorov M., Semenkin E. Acoustic Emotion Recognition: Two Ways of Features Selection Based on Self-Adaptive Multi-Objective Genetic Algorithm. *Proceedings of the International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 2014, p. 851–855.

9. Zitzler E., Thiele L. Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *Evolutionary Computation, IEEE Transactions on,* 1999, vol. 3, no. 4, p. 257–271.

10. Sergienko R., Semenkin E. Competitive Cooperation for Strategy Adaptation in Coevolutionary Genetic Algorithm for Constrained Optimization. *IEEE World Congress on Computational Intelligence (WCCI'2010),* Barcelona, Spain, 2010, p. 1626–1631.

11. Daridi F., Kharma N., and Salik, J. Parameterless genetic algorithms: review and innovation. *IEEE Canadian Review*, 2004, no. (47), p. 19–23.

12. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. F., and Weiss B. A database of german emotional speech. *In Interspeech,* 2005, p. 1517–1520.

13. Grimm M., Kroschel K., and Narayanan S. The vera am mittag German audio-visual emotional speech database. *In Multimedia and Expo, IEEE International Conference on*, IEEE, 2008, p. 865–868.

14. Mori H., Satake T., Nakamura M., and Kasuya H. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 2011, 53 p.

15. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, 2009, vol. 11, Iss. 1.

16. Goutte C., Gaussier E. A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. *ECIR'05 Proceedings of the 27th European conference on Advances in Information Retrieval Research*, 2005, p. 345–359.

17. Popov E. A., Semenkina M. E., Lipinskiy L. V. [Decision making with intelligent information technology ensemble]. *Vestnik SibGAU*. 2012, no. 5 (45), p. 95–99 (In Russ).

Библиографические ссылки

1. Sidorov M., Ultes S., Schmitt A. Emotions are a personal thing: Towards speaker-adaptive emotion recognition // ICASSP. 2014. P. 4803–4807.

2. Eyben F., Wöllmer M., Schuller B. Opensmile: the munich versatile and fast opensource audio feature extractor // Proceedings of the Intern. Conf. on Multimedia. ACM. 2010. P. 1459–1462.

3. Boersma P. Praat, a system for doing phonetics by computer // Glot international. 2002. № 5(9/10). P. 341–345.

4. Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm / M. Sidorov [et al.] // LREC. 2014. P. 3481–3485.

5. Self-adaptive multi-objective genetic algorithms for feature selection / C. Brester [et al.] // Proceedings of Intern. Conf. on Engineering and Applied Sciences Optimization (OPT-i'14). 2014. P. 1838–1846.

6. Kohavi R., John G. H. Wrappers for feature subset selection // Artificial Intelligence. 1997. 97. P. 273–324.

7. Venkatadri M., Srinivasa Rao K. A multiobjective genetic algorithm for feature selection in data mining // International J. of Computer Science and Information Technologies. 2010. Vol. 1, no. 5. P. 443–448.

8. Brester C., Sidorov M., Semenkin E. Acoustic Emotion Recognition: Two Ways of Features Selection Based on Self-Adaptive Multi-Objective Genetic Algorithm // Proceedings of the Intern. Conf. on Informatics in Control, Automation and Robotics (ICINCO). 2014. P. 851–855.

9. Zitzler E., Thiele L. Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach // Evolutionary Computation, IEEE Transactions on. 1999. Vol. 3, no. 4. P. 257–271.

10. Sergienko R., Semenkin E. Competitive Cooperation for Strategy Adaptation in Coevolutionary Genetic Algorithm for Constrained Optimization // IEEE World Congress on Computational Intelligence (WCCI'2010). Barcelona, 2010. P. 1626–1631.

11. Daridi F., Kharma N., Salik J. Parameterless genetic algorithms: review and innovation // IEEE Canadian Review. 2004. № 47. P. 19–23.

12. A database of german emotional speech / F. Burkhardt [et al.] // In Interspeech. 2005. P. 1517–1520.

13. Grimm M., Kroschel K., Narayanan S. The vera am mittag german audio-visual emotional speech database // In Multimedia and Expo, IEEE Intern. Conf. on, IEEE. 2008. P. 865–868.

14. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics / H. Mori [et al.] // Speech Communication. 2011. 53 p.

15. The WEKA Data Mining Software: An Update, SIGKDD Explorations / M. Hall [et al.]. 2009. Vol. 11, iss. 1.

16. Goutte C., Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation // ECIR'05 Proceedings of the 27th European conference on Advances in Information Retrieval Research. 2005. P. 345–359.

17. Попов Е. А., Семенкина М. Е., Липинский Л. В. Принятие решений коллективом интеллектуальных информационных технологий // Вестник СибГАУ. 2012. № 5 (45). С. 95–99.