Multicriteria Neural Network Design in the Speech-based Emotion Recognition Problem

Christina Brester[1], Eugene Semenkin[2], Maxim Sidorov[3], Olga Semenkina[4]
[1,2,4] Institute of Computer Sciences and Telecommunication, Siberian State Aerospace University, Krasnoyarsk, Russia
[3] Institute of Communications Engineering, University of Ulm, Germany
[1]christina.brester@gmail.com, [2]eugenesemenkin@yandex.ru, [3]maxim.sidorov@uni-ulm.de, [4]semenkina.olga@mail.ru

Keywords: Neural network, multicriteria design, cooperative genetic algorithm, speech-based emotion recognition, feature selection.

Abstract: In this paper we introduce the two-criterion optimization model to design multilayer perceptrons taking into account two objectives, which are the classification accuracy and computational complexity. Using this technique, it is possible to simplify the structure of neural network classifiers and at the same time to keep high classification accuracy. The main benefits of the approach proposed are related to the automatic choice of activation functions, the possibility of generating the ensemble of classifiers, and the embedded feature selection procedure. The cooperative multi-objective genetic algorithm is used as an optimizer to determine the Pareto set approximation in the two-criterion problem. The effectiveness of this approach is investigated on the speech-based emotion recognition problem. According to the results obtained, the usage of the proposed technique might lead to the generation of classifiers comprised by fewer neurons in the input and hidden layers, in contrast to conventional models, and to an increase in the emotion recognition accuracy by up to a 4.25% relative improvement due to the application of the ensemble of classifiers.

1 INTRODUCTION

The sphere of human-machine interactions is closely related to *affective computing* (Picard, 1995), which is the interdisciplinary domain including algorithms, systems and devices aimed at recognizing, processing, and simulating human emotions. In most cases all these techniques engage video or audio data to analyse users' emotions. Also there are multi-modal systems which fuse visual information and acoustic characteristics extracted from speech signals. However, in this paper we consider the human emotion recognition problem in the framework of intellectual spoken dialogue systems and, therefore, we apply only audio data.

Previously it was found that compared with various classification models neural networks showed rather high effectiveness for the speech-based emotion recognition problem (Brester *et al.*, 2014). In the experiments conducted a multilayer perceptron (MLP) with one hidden layer trained by the error backpropagation algorithm (BP) was used.

Conventionally, the number of neurons in the hidden layer is proportional to the amount of classes in the sample and the dimensionality of the feature vector. In the case of the emotion recognition problem the quantity of input attributes is very large: generally, we extract 384 acoustic characteristics from the speech signal. As a result, the MLP structure is exaggerated and contains too many neurons in its hidden layer. Moreover, while designing MLPs, researchers have to choose the activation function for each neuron, which is not a trivial task. By default a sigmoid is widely used, despite the fact that there are a lot of other activation functions which are easier in the sense of computational complexity and at the same time might be effectively applied without detriment to the recognition accuracy.

Taking into account these points, we decided to improve the MLP performance by optimizing its structure. In this study we propose a two-objective optimization model which allows us to generate appropriate MLPs based on two criteria: the *classification accuracy* and *computational complexity*. Using this strategy, it is possible to design the MLP whose performance is comparable with the accuracy of the conventional model and whose structure is optimal in the sense of computational complexity. The main advantages of the approach proposed also include the automatic choice of activation functions, the embedded feature selection procedure, and the option of generating the ensemble of classifiers.

The rest of the paper is organized as follows: in Section II a description of the two-objective model for the neural network design and the cooperative genetic algorithm, which is applied to optimize the criteria introduced, are presented. In Section III there is a definition of the speech-based emotion recognition problem and the corpora used. The experiments conducted, the results obtained, and the main inferences are included in Section IV. The conclusions and future work are presented in Section V.

2 PROPOSED APPROACH

2.1 Multicriteria Optimization Model for Neural Network Design

In this study we propose the two-criterion optimization model for neural network design, specifically, for the automatic generation of MLPs with one hidden layer. By taking into account two objectives, it is possible to attain a trade-off between the classification accuracy and computational complexity.

Criterion 1. The relative classification error:

$$minimize\text{: } K1 = E = \frac{N_{incorrectly}}{N_{all}}, \qquad (1)$$

where $N_{incorrectly}$ is the number of instances classified incorrectly, N_{all} is the common number of instances.

Criterion 2. Computational complexity:

$$minimize\text{: } K2 = N_{weights} + \sum_{j=1}^{N_{neurons}} K_j(i), \qquad (2)$$

where $K_j(i) = \frac{T_i^{act}}{T^{weight}}$ is the coefficient reflecting the relative computational complexity of evaluating the *i*-th activation function of the *j*-th neuron; *i* is the identification number of the activation function in the finite set comprised by alternative variants of activation functions; T_i^{act} is the time spent on evaluating the *i*-th activation function; T^{weight} is the time required to process one connection; $N_{weights}$ is the number of connections in the MLP; $N_{neurons}$ is the number of neurons in the MLP. T_i^{act} and T^{weight} are assessed empirically. It is essential to note that K_i is independent of the software used because T_i^{act} is normalized by T^{weight}.

To solve this two-criterion problem, we suggest applying a multi-objective genetic algorithm (MOGA), which operates with binary strings coding diverse MLP structures. Each candidate solution, called a *chromosome*, contains identification numbers of all neurons from the hidden layer (Figure 1). Zero corresponds to the absence of neurons. Input parameters include the set of activation functions with their ID-numbers and the maximum number of neurons in the hidden layer.

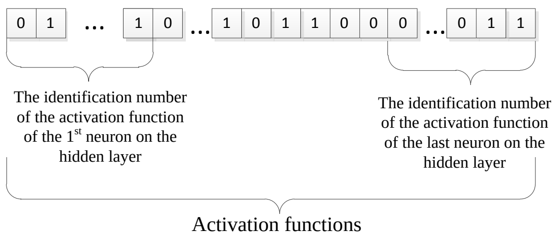

Figure 1: The presentation of the MLP structure as a binary string.

The backpropagation algorithm is applied to train MLPs with different numbers of neurons in the hidden layer and estimate the criterion $K1$.

Moreover, we propose using the cooperative multicriteria genetic algorithm as a multi-objective optimization procedure to diminish the drawbacks of the evolutionary search (Brester *et al.*, 2015a). The next section contains a concise description of this heuristic multi-agent procedure and its advantages.

2.2 Cooperative Multi-objective Heuristic Procedure

While designing a MOGA, researchers are faced with some issues which are related to fitness assignment strategies, diversity preservation techniques, and ways of elitism implementation (Zitzler *et al.*, 2004). To eliminate a number of problems which arise while designing multicriteria evolutionary methods, in this study we use a cooperation of several genetic algorithms (GA) based on various heuristic mechanisms. An island model is applied to involve a few GAs which realize different concepts. Moreover, this model allows us to parallelize calculations and, consequently, to reduce computational time.

Table 1: Basic features of the MOGA used.

MOGA	Fitness Assignment	Diversity Preservation	Elitism
NSGA-II	Pareto-dominance (*niching mechanism*) and diversity estimation (*crowding distance*)	Crowding distance	Combination of the previous population and the offspring
PICEA-g	Pareto-dominance (*with generating goal vectors*)	Nearest neighbour technique	The archive set and combination of the previous population and the offspring
SPEA2	Pareto-dominance (*niching mechanism*) and density estimation (*the distance to the k-th nearest neighbour in the objective space*)	Nearest neighbour technique	The archive set

Generally speaking, an island model (Whitley *et al*., 1997) of a GA implies the parallel work of several algorithms. A parallel implementation of GAs has shown not just an ability to preserve genetic diversity, since each island can potentially follow a different search trajectory, but also could be applied to separable problems. The initial number of individuals M is spread across L subpopulations: $M_i=M/L, \quad i=1,\dots,L$. At each T-th generation algorithms exchange the best solutions (*migration*). There are two parameters: *migration size*, the number of candidates for migration, and *migration interval*, the number of generations between migrations. Moreover, it is necessary to define the island model topology, in other words, the scheme of migration. We use the fully connected topology that means each algorithm shares its best solutions with all other algorithms included in the island model. The multi-agent model is expected to preserve a higher level of genetic diversity. The benefits of the particular algorithm could be advantageous in different stages of optimization.

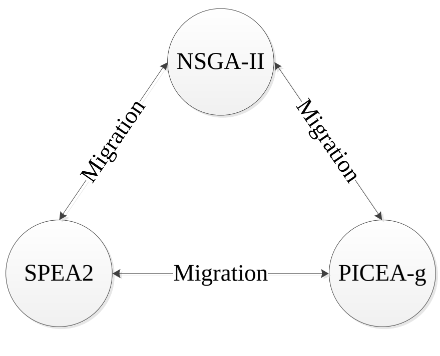

Figure 2: The island model implemented.

In our implementation the *Non-Sorting Genetic Algorithm II* (NSGA-II) (Deb *et al*., 2002), the *Preference-Inspired Co-Evolutionary Algorithm with goal vectors* (PICEA-g) (Wang, 2013), and the *Strength Pareto Evolutionary Algorithm 2* (SPEA2) (Zitzler *et al*., 2002) are used to be involved as parallel working islands (Figure 2).

This multi-agent heuristic procedure does not require additional experiments to expose the most appropriate algorithm for the problem considered. Its performance was thoroughly investigated on the set of test functions CEC2009 (Zhang *et al.*, 2008). The results obtained demonstrated the high effectiveness of the cooperative algorithm and, therefore, we decided to apply it as an optimizer in the neural network design problem.

3 SPEECH-BASED EMOTION RECOGNITION

3.1 Problem Definition

While communicating with humans, machines should perceive the qualities of the user (as people usually do) such as age, gender and emotions to adapt its answers for the particular speaker. *Speech-based emotion recognition* is one of the most essential aspects of the personalization process. Generally, any approach used to solve this recognition problem consists of three main stages.

At first, it is necessary to extract acoustic characteristics from the collected utterances. At the «INTERSPEECH 2009 Emotion Challenge» an appropriate set of acoustic characteristics

Table 2: Statistical description of the used corpora.

Database	Language	Full length (min.)	Number of emotions	File level duration		Notes
				Mean (sec.)	Std. (sec.)	
Emo-DB	German	24.7	7	2.7	1.02	Acted
SAVEE	English	30.7	7	3.8	1.07	Acted
LEGO	English	118.2	3	1.6	1.4	Non-acted
UUDB	Japanese	113.4	4	1.4	1.7	Non-acted

representing any kind of speech signal was introduced. This set of features comprised attributes such as power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants. The mean, minimum, maximum, range and deviation of the following features have also been used: pitch, intensity and harmonicity. The number of characteristics is 384. To get the conventional feature set introduced at INTERSPEECH 2009, the Praat (Boersma, 2002), or OpenSMILE (Eyben *et al*., 2010) systems might be used. Secondly, all extracted attributes or the most relevant of them should be involved in the supervised learning process to adjust a classifier. At the final stage, the signal that has to be analysed is transformed into an unlabelled feature vector (also with the usage of the Praat or OpenSMILE systems) and then the trained classification model receives it as the input data to make a prediction.

3.2 Corpora Description

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

The *Emo-DB* emotional database (German) (Burkhardt *et al.*, 2005) 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 *SAVEE* (Surrey Audio-Visual Expressed Emotion) corpus (English) (Haq *et al*., 2010) was recorded as part of an investigation into audio-visual emotion classification from four native English male speakers. The emotional label for each utterance is one of the standard set of emotions (anger, disgust, fear, happiness, sadness, surprise and neutral).

The *LEGO* emotion database (English) (Schmitt *et al*., 2012) comprises non-acted American English utterances extracted from an automated bus information system of the Carnegie Mellon University at Pittsburgh, USA. The utterances are requests to the Interactive Voice Response system spoken by real users with real concerns. Each utterance is annotated with one of the following emotional labels: angry, slightly angry, very angry, neutral, friendly, and non-speech (critical noisy recordings or just silence). In this study different ranges of anger have been merged into a single class and friendly utterances have been deleted. This pre-processing results in a 3-class emotion classification task.

The *UUDB* (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database (Japanese) (Mori *et al*., 2011) 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 the pleasantness and arousal axes are used. 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.

There is a statistical description of the used corpora in Table II.

4 EXPERIMENTS AND RESULTS

To investigate the effectiveness of the approach proposed, we performed several experiments. Firstly, the conventional MLP classifier implemented in the program system *WEKA* (Hall *et al*., 2009) was applied. This model was trained with the BP algorithm and contained one hidden layer with *[(NumberOfFeatures+NumberOfClasses)/2+1]* –neurons (the number of features was equal to 384, the number of classes varied from 3 to 7 depending on the database used). The activation function for all of the neurons was a sigmoid. For each database the 6-fold cross-validation procedure (and also in the next experiments) was conducted to assess the averaged *F-score* metric (Goutte *et al*., 2005): the more effective the classifier applied, the higher the F-score value obtained.

Then we used the proposed two-criterion optimization model to design MLP classifiers automatically. The set of possible activation functions was defined beforehand (Table 3).

Table 3: Activation functions used.

№	Activation function	K_i value
1	$f(x)=\begin{cases} -1, & x<-1 \\ x, & -1\le x\le 1 \\ 1, & x>1 \end{cases}$	6.46
2	$f(x)=1$	2.69
3	$f(x)=e^{-\frac{x^2}{2}}$	22.48
4	$f(x)=tanh(x)$	23.14
5	$f(x)=\frac{1}{1+e^{-x}}$	22.20
6	$f(x)=1-e^{-\frac{x^2}{2}}$	20.55
7	$f(x)=atan(x)$	27.01

Moreover, to evaluate the second criterion *'computational complexity'*, we estimated K_i coefficients empirically. The time spent on evaluating the *i*-th activation function T_i^{act} was normalized by the T^{weight} value obtained as the time required to multiply two real numbers and then to add another real value to this sum (we simulated the processing of a new connection in the MLP structure). It might be noticed that the K_i values for some of these functions stand out significantly.

In the second experiment the cooperative multi-objective genetic algorithm was used as an optimizer. It is well-known that in contrast to one-criterion GAs, the outcome of MOGAs is the set of non-dominated points which form the Pareto set approximation and, therefore, it is necessary to choose one solution from the set of alternative candidates. In addition to the training sample and the test one, we used the evaluation set, which was 20% of the training sample, to compare non-dominated points based on the classification accuracy on these examples. 80% of the training set were used by the MOGA to assess the first criterion *'the relative classification error'*: to obtain the averaged value of this metric, we conducted the 3-fold cross-validation procedure for each binary string coding the MLP structure. The number of epochs in the BP algorithm was equal to 25. For each component of the MOGA (NSGA-II, PICEA-g, and SPEA2) the following settings were defined: binary tournament selection, uniform recombination and the mutation probability $p_m=1/n$, where n is the length of the chromosome. All islands had an equal amount of resources (20 generations and 30/3 = 10 individuals in populations), the migration size was equal to 3 (in total each island got 6 points from two others), and the migration interval was equal to 5 generations. Finally, the evaluation sample was used to compare the classification accuracy provided by all of the alternative MLP structures and to choose the most effective model based on these values (at this stage the number of epochs in the BP algorithm was equal to 100). Then the training and evaluation instances were merged and used by the BP algorithm (the number of epochs was equal to 250) to find the classification accuracy on the test set.

However, it is possible to take advantage of the set of alternative solutions. We repeated the previous experiment, but we did not choose only one solution amid all of the non-dominated points. Based on the classification accuracy assessed on the evaluation set, 15 different models with the highest performance were defined (the common number of non-dominated alternatives was equal to 3·10+2·7=44, where 10 was the population size for all three islands and 7 was the outer set size for the PICEA-g and SPEA2 algorithms). Then these most effective MLP models were included in the ensemble of classifiers to make a collective decision based on the majority rule.

Owing to the binary representation of the MLP structure, the feature selection procedure might be embedded in the model design process (Brester *et al.*, 2015b). The vector of inputs is also presented as a part of the binary string, where *unit* and *zero* correspond to a relative attribute and an irrelative one respectively (Figure 3).

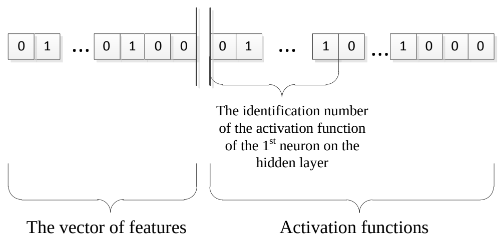

Figure 3: The presentation of the MLP structure with the feature selection procedure.

Next, two previous experiments were repeated with extended binary strings which coded not only the hidden layer of the MLP structure but also the feature vector. Due to the greater number of genes in a chromosome, it was necessary to increase the quantity of computational resources which the MOGA was provided with. In these experiments for all of the islands the number of generations was equal to 30, the population contained 60/3 = 20 individuals. (The migration size was equal to 5 and the migration interval was equal to 5 generations). While generating the initial population, the probability of 1 in the part of the chromosome corresponding to the feature vector was equal to 0.8. Firstly, we chose one most effective MLP structure from the set of non-dominated points and estimated the F-score value. Secondly, we formed the ensemble of classifiers and again defined the F-score metric.

Table 4 contains the results obtained in all of the experiments conducted.

Table 4: Experimental results.

Database	Classifier	F-score	K2 criterion value	The number of neurons in the hidden layer	Dimensionality of the feature vector
Emo-DB	Conventional MLP (WEKA)	80.83	81759.8	197	384
	MLP designed by the MOGA	80.75	62492.6	152.17	384
	Ensemble of MLPs designed by the MOGA	**82.90**	62276.5	151.6	384
	MLP designed by the MOGA with feature selection	79.94	40997.0	139	265
	Ensemble of MLPs designed by the MOGA with feature selection	81.56	41895.7	142.03	268.12
SAVEE	Conventional MLP (WEKA)	59.55	81759.8	197	384
	MLP designed by the MOGA	61.69	62212.1	151.5	384
	Ensemble of MLPs designed by the MOGA	**62.02**	62677.2	152.7	384
	MLP designed by the MOGA with feature selection	60.44	46144.9	155.7	269.2
	Ensemble of MLPs designed by the MOGA with feature selection	61.58	43208.3	149.644	261.5
LEGO	Conventional MLP (WEKA)	68.19	80058.6	195	384
	MLP designed by the MOGA	66.03	62678.2	154.3	384
	Ensemble of MLPs designed by the MOGA	71.05	62172.6	153.1	384
	MLP designed by the MOGA with feature selection	66.81	44910.7	151.7	273.0
	Ensemble of MLPs designed by the MOGA with feature selection	**71.10**	44107.5	149.9	271.7
UUDB	Conventional MLP (WEKA)	49.34	80276.8	195	384
	MLP designed by the MOGA	47.18	62667.9	154	384
	Ensemble of MLPs designed by the MOGA	50.68	63033.0	154.7	384
	MLP designed by the MOGA with feature selection	49.32	45330.0	154.0	272.0
	Ensemble of MLPs designed by the MOGA with feature selection	**51.34**	43603.4	148.844	269.178

The rows *'Ensemble of MLPs designed by the MOGA'* and *'Ensemble of MLPs designed by the MOGA with feature selection'* contain the averaged results for MLP structures in the ensemble.

We analysed the experimental results statistically: a t-test with the significance level p=0.05 exposed that the conventional MLP did not outperform any of models (or any ensemble of models) designed by the cooperative MOGA in the sense of classification performance.

Due to the usage of the two-criterion optimization model, we managed to find MLP structures which were also effective in terms of computational complexity. They not only had fewer neurons in the hidden layer, but also the activation functions required less computational time. These classifiers might be especially effective if it is necessary to make predictions in real time.

At the same time we increased the F-score values significantly with the application of MLP ensembles. For all of the databases MLP ensembles with or without the feature selection procedure demonstrated the best results. The relative improvement of F-score values compared with the effectiveness of the conventional MLP was equal to: 2.56% for *Emo-DB*, 4.15% for *SAVEE*, 4.25% for *LEGO*, and 4.05% for *UUDB*.

Moreover, it is important to note that the usage of the embedded feature selection procedure allowed us to simplify MLP structures and decrease their computational complexity significantly.

5 CONCLUSION

In this paper we proposed the two-criterion optimization model to design MLP classifiers automatically for the speech-based emotion recognition problem. The main benefit of this approach is the opportunity to generate effective MLP structures taking into consideration two objectives *'classification performance'* and *'computational complexity'*.

In the experiments conducted it was revealed that this technique allowed us to design MLP classifiers with simpler structures, whose accuracy was comparable with (or even higher than) the performance of conventional MLPs containing more neurons in the hidden layer.

In the framework of this technique, it is also possible to design ensembles of classifiers; their application leads to the essential improvement of the classification quality.

The binary representation of MLP structures permitted us to embed the feature selection procedure and additionally to simplify classifiers.

Finally, there are some other questions related to the human-machine communication sphere. The proposed scheme might be applied without any changes to the speech-based speaker identification problem as well as to speaker gender or age recognition.

ACKNOWLEDGEMENTS

Research is performed with the financial support of the Ministry of Education and Science of the Russian Federation within the federal R&D programme (project RFMEFI57414X0037).

REFERENCES

Boersma, P., 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10), pp. 341–345.

Brester, Ch., Sidorov, M., Semenkin, E., 2014. Speech-based emotion recognition: Application of collective decision making concepts. *Proceedings of the 2nd International Conference on Computer Science and Artificial Intelligence (ICCSAI2014)*, pp. 216-220.

Brester, Ch., Semenkin, E., 2015a. Cooperative multi-objective genetic algorithm with parallel implementation. *ICSI-CCI 2015*, Part I, LNCS 9140, pp. 471–478.

Brester, Ch., Semenkin, E., Sidorov, M., Kovalev, I., Zelenkov, P., 2015b. Evolutionary feature selection for emotion recognition in multilingual speech analysis // *Proceedings of IEEE Congress on Evolutionary Computation (CEC2015)*, pp. 2406–2411.

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

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6 (2), pp. 182-197.

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

Goutte, C., Gaussier, E. 2005. 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*, pp. 345–359.

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

Haq, S., Jackson, P., 2010. Machine Audition: Principles, Algorithms and Systems, chapter Multimodal Emotion Recognition. IGI Global, Hershey PA, pp. 398–423.

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

Picard, R.W., 1995. Affective computing. Tech. Rep. Perceptual Computing Section Technical Report No. 321, MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139.

Schmitt, A., Ultes, S., and Minker, W., 2012. A parameterized and annotated corpus of the cmu let’s go bus information system. *Proceedings of International Conference on Language Resources and Evaluation (LREC*), Istanbul, Turkey, pp. 3369-3373.

Wang, R., 2013. Preference-Inspired Co-evolutionary Algorithms. A thesis submitted in partial fulfillment for the degree of the Doctor of Philosophy, University of Sheffield.

Whitley, D., Rana, S., and Heckendorn, R., 1997. Island model genetic algorithms and linearly separable problems. *Proceedings of AISB Workshop on Evolutionary Computation*, vol.1305 of LNCS, pp. 109–125.

Zhang, Q., Zhou, A., Zhao, S., Suganthan, P. N., Liu, W., Tiwari, S., 2008. Multi-objective optimization test instances for the CEC 2009 special session and competition. University of Essex and Nanyang Technological University, Tech. Rep. CES-487, 2008.

Zitzler, E., Laumanns, M., Bleuler, S., 2004. A Tutorial on Evolutionary Multiobjective Optimization. In: Gandibleux X., (eds.): *Metaheuristics for Multiobjective Optimisation. Lecture Notes in Economics and Mathematical Systems*, vol. 535, Springer.

Zitzler, E., Laumanns, M., Thiele, L., 2002. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. *Evolutionary Methods for Design Optimisation and Control with Application to Industrial Problems EUROGEN* 2001 3242 (103), pp. 95–100.