OPT-*i*
An International Conference on
Engineering and Applied Sciences Optimization
M. Papadrakakis, M.G. Karlaftis, N.D. Lagaros (eds.)
Kos Island, Greece, 4-6 June 2014

SELF-ADAPTIVE MULTI-OBJECTIVE GENETIC ALGORITHMS FOR FEATURE SELECTION[1]

C. Brester[2], E. Semenkin[2], M. Sidorov[3], W. Minker[3]

[2] Siberian State Aerospace University,
Krasnoyarsky rabochy avenue, 31, 660014, Krasnoyarsk, Russia,
[2]abahachy@mail.ru, [2]eugenesemenkin@yandex.ru

[3] University of Ulm,
Albert-Einstein-Allee, 43, 89081 Ulm, Germany,
[3]{ maxim.sidorov, wolfgang.minker}@uni-ulm.de

Keywords: Multi-Objective Optimization, Genetic Algorithm, Self-Adaptation, Feature Selection, Emotion Recognition.

Abstract. *Incorporating multi-objective optimization procedures in the classification process allows to achieve a trade-off between the accuracy of the final solution and the number of features involved in supervised learning. In our research we investigate a number of self-adaptive multi-objective genetic algorithms as a tool to select the most essential attributes from the database. A probabilistic neural network is implemented to evaluate the relevancy of reduced feature sets. The high performance of the developed approach is demonstrated on the speech-based emotion recognition problem. For the engaged data set it became possible not only to improve the classification accuracy by up to 28.83% but also to decrease the number of features from 384 to 85.5 on average.*

[1] Research is fulfilled with the support of the Russian Science Foundation.

1 INTRODUCTION

Classification procedures based on the supervised learning approach imply the presence of labeled sampling data. In most cases we have to deal with raw information which includes highly-correlated features, measures with errors or attributes with the low variation level. As a result, using of irrelevant data by the learning algorithm is likely to deteriorate its performance.

Apparently, improvement of the classifier's predictive ability might be achieved through eliminating non-informative features from the database. In this paper we propose to accomplish the heuristic search of essential attributes by means of multi-objective genetic algorithms (MOGAs). The possibility to take into account several criteria at once allows to minimize the relative classification error and the number of selected features simultaneously.

To incorporate MOGAs in the developed scheme it was necessary to elaborate some significant aspects. First, the self-adaptation concept [1] was borrowed to provide the realized approach with automatically adjusted items. Actually, it was a good alternative to occasional choice of genetic operator variants or multiple runs of the algorithm to reveal the best settings. Secondly, applied MOGAs should be oriented to the nature of optimized criteria because algorithms had to operate with both continuous and discrete objective functions in diverse value areas.

The proposed approach was applied to the speech-based emotion recognition problem that was one of the crucial opportunities to improve the quality of spoken dialogue systems. Two data sets (37- and 384-dimensional feature vectors) representing the acted German language corpus were engaged in the series of experiments for comparison the effectiveness of different MOGAs. It was found that due to implementation of the heuristic search there was an opportunity to improve the emotion recognition accuracy by up to 28.83% and reduce the number of features from 384 to 85.5 on average.

This paper is comprised of the following parts: Section 2 provides the brief overview of important studies related to application of heuristic methods in the feature selection process. Section 3 includes the description of used MOGAs, their basic stages and modifications. The problem definition, conducted experiments and results are introduced in Section 4. Conclusion and future plans are presented in Section 5.

2 RELATED WORKS

Generally, the feature selection procedure can be organized as the *wrapper* approach or the *filter* one [2]. The first technique involves classification models to evaluate the relevancy of each feature subset. Although it requires high computational resources, this approach demonstrates the effective work due to adjustment to an applied classifier. The second technique is referred to the preprocessing stage because it extracts information from the data set and reduces the number of attributes taking into consideration such measures as consistency, dependency, and distance. On the one hand, this approach needs significantly fewer calculations therefore it is rather effective in the computational effort sense. However, it does not co-operate with a learning algorithm during feature selection and so ignores its performance entirely.

Yang and Hanovar (1998) used one-criterion genetic algorithm (GA) to determine relevant attributes in order to improve quality of classification realized with neural networks [3]. Li Zhuo *et al.* (2008) accomplished classification of hyperspectral images with support vector machine, they also engaged one-criterion GA to remove non-informative features [4]. In both cases the feature selection procedure was combined with supervised learning algorithms based on the wrapper approach scheme.

Lanzi (1997) offered to apply a heuristic method to extract attributes before executing classification [5]. The inconsistency rate was used by GA to assess the relevancy of reduced data sets. Due to implementation of the filter approach it became possible not only to achieve the high performance of C4.5 inductive algorithm but also to lower a computational cost.

Development of multi-objective optimization algorithms allowed to embed them in the feature selection procedure to take into account several criteria. Venkatadri and Srinivasa (2010) introduced a set of measures such as *Attribute Class Correlation, Inter- and Intra-Class Distances, Laplasian Score, Representation Entropy* and *the Inconsistent Example Pair measure* to estimate the quality of reduced databases. They investigated various combinations of these criteria by means of the Non-dominated Sorting Genetic Algorithm (NSGA-II) [6]. Hamdani *et al.* (2007) also implemented NSGA-II to attain a compromise between the number of extracted attributes and the classification accuracy evaluated with 1-NN classifier [7]. These are examples of MOGA realization in the framework of the filter and the wrapper approach respectively.

3 MOGA IN FEATURE SELECTION

Three self-adaptive MOGA were developed to incorporate them into the wrapper feature selection approach. In all cases two criteria were introduced: the first one was the number of selected features and the second one was the relative classification error, both of them were minimized. We used binary representation to code informative and non-informative attributes (*unit* and *zero* correspondingly). Probabilistic neural networks (PNN) [8] were engaged as a classifier to estimate the relevancy of selected feature subsets.

This section provides the description of realized self-adaptive MOGAs.

3.1 Preference-inspired co-evolutionary algorithm using goal vectors

Preference-inspired co-evolutionary algorithm using goal vectors (PICEA-g) proposed by Wang (2013) [9] includes the following steps:

1. Generate an initial population and evaluate objective values for individuals. Find non-dominated candidate solutions in the population and copy them into the *archive*. Determine the set of *goal vectors* as a number of targets randomly generated within the goal vector bounds.
2. Produce the offspring solutions with *genetic operators*: selection, crossover and mutation. Evaluate objective values for new generated individuals.
3. Pool together parents and children; compile the common set of objective values.
4. Append to the set of goal vectors the additional targets generated within the determined bounds.
5. Assign fitness values for goal vectors and for individuals in the united population.
6. Form the new population and the set of goal vectors based on their fitness.
7. Update the archive with new non-dominated solutions.
8. Check the stopping criterion: if it is satisfied then finish the search with the archive set, otherwise proceed from the second step.

In Steps 1 and 4 decision maker preferences are incorporated into the algorithm using goal vectors. They represent points generated in the criteria search space within the bounds which are determined according to the rule:

$$\begin{gathered} g_i^{min} = min(BestF_i), \\ \Delta F_i = max(BestF_i) - min(BestF_i), \\ g_i^{max} = min(BestF_i) + \alpha \times \Delta F_i , \end{gathered} \tag{1}$$

where g_i^{min} is the lower bound and g_i^{max} is the upper one for the i-th goal vector component, $BestF_i$ is the best value of the i-th objective function amid solutions in the archive, $i = 1,...,M$, M is the number of criteria.

The recommended value of the α parameter is 1.2. However, there is an opportunity to adjust it in order to concentrate the search on the certain criterion. While investigating we discovered that it was much easier for the algorithm to decrease the amount of features than to minimize the relative classification error. Therefore we assigned $\alpha = 0.6$ for the first criterion to narrow the corresponding bounds. This strategy allowed to moderate the feature reduction and to stimulate the classification accuracy improvement.

In Step 2 two conventional genetic operators were implemented: tournament selection and uniform recombination. For realizing the mutation operator the self-adaptive scheme proposed by Daridi *et al.* (2004) [10] was engaged. This heuristics is equal to:

$$p_m = \frac{1}{240} + \frac{0.11375}{2^t}, \tag{2}$$

where p_m is the mutation probability, t is the current generation number.

3.2 Multi-objective evolutionary algorithm based on decomposition

The self-adaptive version of the multi-objective evolutionary algorithm based on decomposition with dynamic resource allocation (MOEA/D-DRA) proposed by Zhang *et al*. (2009) [11] also was involved as the optimizer in the feature selection procedure. MOEA/D-DRA was one of the leaders in CEC 2009 MOEA competition [12] on the set of unconstrained problems with real variables. We adapted this scheme to binary representation using conventional genetic operators: selection (proportional, tournament and rank), crossover (one-point, two-point and uniform) and mutation (weak, average and strong).

There is a brief description of basic steps in MOEA/D-DRA:

1. *Initialization.* Generate the initial population and the set of weight vectors.
2. *Selection of subproblems.* Using utilities of subproblems select a number of them for the search.
3. For each selected subproblem do*:*

3.3 Generate the intermediate set of individual indexes which participate in producing the offspring.

3.4 *New solution generation*. Select the parents from the set formed in the previous step. Fulfill crossover and mutation.

3.5 *Upgrade the solutions*. Compare the effectiveness of the new obtained individual and some other solutions which indexes are contained in the intermediate set. Replace the solutions with the new candidate if their effectiveness is lower.

4. *Stopping criterion.* If it is satisfied then finish the search with the current population.
5. *Upgrade the subproblem utilities.* Based on improvement of objective functions recalculate the subproblem utilities. Go to Step 2.

In Step 3.4 one of the genetic operator variants should be applied. The self-adaptation mechanism for automatic choosing the appropriate operator type was realized. We introduced application probabilities q_i^k for each *i*-th variant of the *k*-th operator. In the beginning all variants of every genetic operator had equal probabilities $q_i^k = {1}/{n^k}$, n^k is the number of different variants of a certain genetic operator, $i, k = \overline{1,3}$.

After the adaptation interval (that was the certain number of objective function evaluations) the probabilities were recalculated taking into account fitness of individuals generated by the given operator. The main idea was borrowed from Banzhaf's article [13] with the only difference: the variable named « $ratio_i^k$ » was a fitness sum of individuals generated with the *i*-th variant of the *k*-th operator.

Below there is a rule for probability q_i^k calculation:

$$q_i^k = \frac{0.2}{n^k} + 0.8 \cdot \frac{ratio_i^k}{scale^k}, \qquad (3)$$

where $scale^k = \sum_i ratio_i^k$. The first summand does not allow any probability to be equal to zero (that makes all variants of operators available throughout the algorithm execution).

In MOEA/D-DRA fitness values of solutions are determined based on the rule (*Tchebycheff approach*):

$$g(\bar{x} \mid \bar{\lambda}, \bar{z}^*) = max_{1<i<M}\{\lambda_i \mid f_i(\bar{x}) - z_i^* \mid\}, \qquad (4)$$

where $\bar{z}^* = (z_1^*,...,z_M^*)^T$ is the reference point and in the case of feature selection it is $\bar{z}^* = (0,0)^T$; $\bar{\lambda}$ is a weight vector.

However, in the feature selection procedure we deal with two criteria which possess values from different intervals. The relative classification error varies from 0 to 1, whereas the number of features might be equal to several hundred or even thousand. Therefore it is impossible to compare absolute differences between the objective value and the reference point of these criteria. As an alternative the relative improvement of each objective function might be used:

$$g(\bar{x} \mid \bar{\lambda}, \bar{z}^*) = max_{1<i<M}\{\lambda_i \left| \frac{f_i(\bar{x}) - z_i^*}{\Delta f_i} \right|\}, \qquad (5)$$

where Δf_i is the range space of the *i*-th criterion.

In addition, the relative classification error is more essential criterion and there is an opportunity to concentrate the search in promising regions using weight vectors. In fact we are able to generate them in the special way to obtain the particular part of the Pareto frontier approximation and not to spend computational resources on the whole front in vain. Therefore in the feature selection procedure weight vectors were uniformly generated within the bounds: $0 \le \lambda_1 \le 0.5,\ 0.5 \le \lambda_2 \le 1$ and $\lambda_1 + \lambda_2 = 1$.

3.3 Genetic algorithm with the rank aggregating fitness function (GA-RAFF)

One of the prime approaches to solving multi-objective optimization problems is the usage of aggregating criteria. In spite of its simplicity we borrowed this concept to adapt it for feature selection. Our method is based on the conventional single-objective GA (Fig. 1) [14] where the fitness function includes criterion values as ranks with weight coefficients evolving through the algorithm execution.

In contrast to other MOGAs which operate with the set of non-dominated solutions, GA-RAFF returns just a single point and a decision maker should not choose the final solution among obtained alternatives.

The next scheme illustrates the fitness assignment process:

1. Compute the values of objective functions for all individuals.

2. Rank individuals based on criteria values: produce sorted arrays for each objective; associate the worst and the best individuals with *1* and *N* respectively, where *N* is the population size; average ranks for solutions with equal objective values.
3. Aggregate the individual ranks of all criteria using weight coefficients.

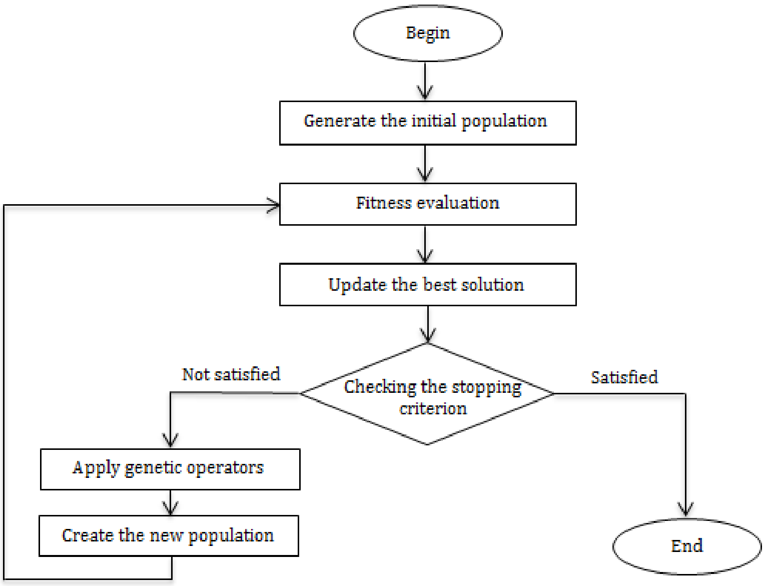

Figure 1: Single-objective GA

Objective coefficients in the aggregating sum are also changed through generations. In the beginning weights are equal: $w_1 = 0.5$ and $w_2 = 0.5$, $w_1 + w_2 = 1$, where w_1 corresponds to the relative classification error and w_2 relates to the number of selected features. For several generations criteria have the same significance, however, in order to direct the search toward minimizing the classification error weights are recalculated:

$$\begin{array}{ll} \textit{if } t \textit{ is a multiple of } T_{adapt}, & \textit{then} \quad w_1 = w_1 - \frac{0.5 \cdot T_{adapt}}{T}, w_2 = w_2 + \frac{0.5 \cdot T_{adapt}}{T}, \\ & \textit{otherwise} \quad w_1 = w_1, w_2 = w_2 \end{array} \quad (6)$$

where t is the current generation number, T is the total number of generations, T_{adapt} is the adaptation interval. This technique allows to control the feature elimination and supports the tendency to find more precise solutions.

The algorithm uses the same adaptation concept for genetic operator adjustment as it is described in Section 3.2.

4 PERFORMANCE ASSESSMENT

The developed approach might be worthwhile in various applications. In particular, this paper proves the reasonableness of using this technique for improvement of the "human-machine" communication. While spoken dialogue systems are quite good at recognizing human speech, there are some open questions.

One of the current trends is an aspiration to teach machines to reveal human personal characteristics like educational level, gender or age to adapt their answers for the user as people usually do during the conversation. Moreover, state-of-the-art systems are able to identify human emotions. But existing recognition methods based on the supervised learning approach have to cope with a huge amount of features which are extracted from voice records. Therefore it is utterly important to determine the relevant feature sets in order to decrease the quantity of engaged data and improve the learning algorithm performance.

4.1 Problem description

In this study the speech-based emotion recognition problem is presented with the Berlin database [15] provided by the Technical University of Berlin. It includes 535 emotive records spoken by 10 actors (male and female) simulating one the following senses: neutral, anger, fear, joy, sadness, boredom or disgust.

Baseline and extended data sets comprised of 37- and 384-dimensional feature vectors were involved in experiments. Acoustic characteristics of each waveform were extracted with the Praat [16] and the openSMILE [17] systems. For instance, the 37-dimensional vector includes average values of power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants and also mean, minimum, maximum, range and deviation of such features as pitch, intensity and harmonicity.

4.2 Experiments and results

To evaluate the effectiveness of the developed approach we compared the PNN-classifier performance on the full set of features (Table 1) and the MOGA-PNN system execution on the reduced set of attributes (Table 2).

Data set	Classification accuracy	Number of features
Baseline	56.68	37
Extended	58.90	384

Table 1: Classification results on full feature sets

For every experiment the classification procedure was run 20 times. The data set was randomly divided into training and test samples in proportion 70-30%. To estimate the relevancy of solutions which MOGAs operated, the validation sample was generated as 20% of the training data set.

In addition, it was noticed that the PNN performance depends essentially on the sample division. Therefore, we produced the averaged estimations of the relative classification error on the validation sample for all candidate solutions through multiple running (15 times) with the random data division. The test set was used to obtain the final assessment of the model's predictive ability.

№	Feature selection procedure	Classification accuracy	Average number of features	Gain
1	PICEA-g	73.05	85.5	28.83%
2	MOEA/D-DRA	69.73	160.1	23.02%
3	GA-RAFF	73.02	101.5	28.83%
4	SPEA	71.46	68.4	26.08%
5	GA	70.70	155.1	24.74%
6	PCA	43.66	129.3	-22.97%

Table 2: The effectiveness of feature selection procedures

In all experiments MOGAs were provided with the equal amount of resources (for each run 10100 candidate solutions were examined in the search space). For methods using the set of non-dominated candidates the final solution was determined as the point from the archive with the lowest error on the validation sample.

Table 2 contains the relative accuracy improvement due to the usage of MOGAs in comparison with the PNN-performance on the baseline data set. In addition to MOGAs described above, this table provides the information about the effectiveness of other alternative self-adaptive methods: the Strength Pareto Evolutionary Algorithm (SPEA) that was investigated in [18] and one-criterion GA minimizing the relative classification error and realizing the same self-adaptation concept as it was mentioned in Section 3.4. Moreover, we compared these heuristic procedures with conventional Principal Component Analysis (PCA) with the 0.95 variance threshold.

4.3 Discussion

According to obtained results, the heuristic search for feature selection in the emotion recognition problem is much more effective than application of the PCA-based technique that leads to decreasing the classification accuracy.

Next, using Wilcoxon nonparametric criteria (with significance level $\alpha = 0.05$) it might be found that the GA-PNN system which is not oriented to the feature reduction does not outperform any approaches taking into consideration two criteria (the number of features and the relative error).

Conducted experiments demonstrate the highest effectiveness of PICEA-g and GA-RAFF algorithms in the developed scheme. Based on Wilcoxon nonparametric criteria ($\alpha = 0.05$) we even may conclude that there is no significant difference between classification accuracy values obtained with using these methods. Although PICEA-g finds solutions with the fewer number of features, in comparison with GA-RAFF, it requires additional computational resources during the clusterization stage or the goal vector fitness assignment. Thus, in spite of its simplicity, GA-RAFF is able to compete with other MOGAs realizing the Pareto dominance idea and the elitism concept.

5 CONCLUSION

This study reveals advantages of using the MOGA-PNN combination in feature selection by the example of the speech-based emotion recognition problem. Obtained results reflect superiority of the developed approach in contrast to application of the PNN-classifier or the PCA-PNN hybrid system. The multi-objective technique allows to find a compromise between the feature reduction and the classification accuracy improvement.

In the conducted experiments the MOGA-PNN system succeeds in the emotion identification: on the one hand, it is able to select approximately 86 the most relevant attributes from 384 ones and, on the other hand, this approach permits to improve the classification accuracy by up to 28.83% (from 56.68% to 73.05%).

Future lines of this study lie in the following directions: first, it is reasonable to incorporate more precise classifiers into the feature selection procedure, secondly, there is an opportunity for the most effective MOGAs to co-operate with each other to achieve better results and, next, this approach should be investigated on the set of other classification problems.

Moreover, realized MOGAs might be used in the framework of the filter approach to fulfill the feature selection as the data preprocessing stage where a number of criteria characterizing consistency, dependency or distance might be optimized.

REFERENCES

[1] A.E. Eiben, R. Hinterding, Z. Michalewicz, Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2): pp. 124–141, 1999.

[2] R. Kohavi, G. H. John, Wrappers for feature subset selection. *Artificial Intelligence*, 97: pp. 273-324, 1997.

[3] J. Yang and V. Hanovar, Feature subset selection using a genetic algorithm. *Journal of IEEE Intelligent Systems*, vol. 13, pp. 44-49, 1998.

[4] L. Zhuo, J. Zheng, F. Wang, X. Li, B. Ai, J. Qian, A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVII, part B7*, pp. 397–402, 2008.

[5] PL. Lanzi, Fast feature selection with genetic algorithms: a filter approach. *IEEE International Conference on Evolutionary Computation*, pp. 537–540, 1997.

[6] M. Venkatadri, K. Srinivasa Rao, A multiobjective genetic algorithm for feature selection in data mining. *International Journal of Computer Science and Information Technologies, vol. 1, no. 5*, pp. 443–448, 2010.

[7] T.M. Hamdani, J. Won, A.M. Alimi, F. Karray, Multi-objective feature selection with NSGA II. In *Adaptive and Natural Computing Algorithms Lecture Notes in Computer Science, vol. 4431*, pp. 240–247, 2007.

[8] D. F. Specht, Probabilistic neural networks. *Neural networks, 3(1):* pp. 109–118, 1990.

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

[10] F. Daridi, N. Kharma, and J. Salik, Parameterless genetic algorithms: review and innovation. *IEEE Canadian Review*, (47): pp. 19–23, 2004.

[11] Q. Zhang, W. Liu, and H Li, The Performance of a New Version of MOEA/D on CEC09 Unconstrained MOP Test Instances. In *CEC'09 Proceedings of the Eleventh conference on Congress on Evolutionary Computation*, pp. 203-208, 2009.

[12] Q. Zhang, A. Zhou, S. Zhao, P. N. Suganthan, W. Liu, and S. Tiwari. 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.

[13] J. Niehaus, W. Banzhaf. Adaption of Operator Probabilities in Genetic Programming, In *Proceedings of the 4th European Conference on Genetic Programming, Lecture Notes In Computer Science, vol. 2038*, Springer-Verlag, Berlin, Heidelberg, pp. 325–336, 2001.

[14] J.H. Holland, Adaptation in natural and artificial systems / J.H. Holland – Ann Arbor. MI: University of Michigan Press, 1975.

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

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

[17] F. Eyben, M. Wöllmer, and B. Schuller, Opensmile:the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pp. 1459–1462, 2010. ACM.

[18] M. Sidorov, C. Brester, W. Minker, E. Semenkin, Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm, In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference,* 2014. – in press.