Scheme of Big-Data Supported Interactive Evolutionary Computation

Guo-sheng HAO¹**, Na GUO¹, Gai-ge WANG¹, Zhao-jun ZHANG² and De-xuan ZOU²

¹School of Computer Science and Technology, Jiangsu Normal University, Xuzhou, 221116, China
²School of Electrical Engineering and Automation, Jiangsu Normal University, Xuzhou, 221116, China

*Corresponding author

Keywords: Interactive evolutionary computation, Big-data, Models repository, Transfer learning.

Abstract. On the one hand, for the limitation that a user in Interactive Evolutionary Computation (IEC) is apt to feel tired during the solutions evaluation, the result of IEC is not so desirable. On the other hand, the many-years and many-users accumulated data of the evaluation in IEC is enormous, which could be a pool of information about users’ preference. By mining the big data of accumulated evaluation information, preference models can be established to form a model repository. When a new coming customer tries to get his/her favorite design aided by IEC, the nearest model in the repository can be selected as the result of model matching and the corresponding optimum can be directly transferred to accelerate the optimization. For this purpose, transfer learning is adopted. The simulated experiments validated this scheme and showed that the performance of big-data supported IEC is more related with volume of the models in big data, but less related with problems’ difficulty.

Introduction

Evolutionary Computation (EC) is problem independent and has been reported to perform relatively well on problems such as: search, optimization, and artificial intelligence [1]. However, in order to obtain the most favorable outputs from interactive systems that create or retrieve graphics or music, the outputs must be subjectively evaluated. It is difficult or even impossible to design explicit functions of human’s evaluation. Interactive Evolutionary Computation (IEC) is an optimization method by incorporating human’s preference as fitness for solutions. Therefore, IEC has been applied to many art designs such as personalized music, popular software menu and so on.

The main problem in IEC is users’ fatigue [2], because a user in IEC is apt to feel tired after evaluating many solutions. Aiming at this problem, many studies to estimate fitness [3] based on knowledge learning have been reported. If fitness can be estimated, the user can be replaced by the machine [4]. For this reason, in IEC, the information transferred from the user to system is generally insufficient for the system to well model the user’s preference. This is shown in Figure 1, where the information about the user’s preference is labeled as \( I_U \) and the information that the system gets from the user at time \( t \) is labels as \( I_M(t, U) \). Then the information that the system has not attained is:

\[
\Delta I_{M,U} (t) = I_U - I_M(t, U).
\] (1)

Therefore, in order to well model the user’s preference and to increase the quality of the optimum, one of methods is to decrease \( \Delta I_{M,U}(t) \). This is the purpose for us to use big data in this paper.

Nowadays, the big data has been applied in many areas, such as education, business and so on, to understand the users better. In the study of IEC, a situation is that accumulated evaluation data is enormous, and they were not jointly utilized. Taking them as a pool of users’ preference information, big data technique can be used to improve the performance of IEC without worsening the main problem in IEC, i.e. the user’s’ fatigue. Therefore, the scheme of big-data supported IEC is put forward in this paper.
Scheme of Big Data Supported Interactive Evolutionary Computation

In the past years, many users and volunteers took part in IEC experiments or applications, and multi-dimensional preference evaluation data, including fuzzy type evaluation [5], accurately numeric type evaluation [6], rank type evaluation [7], eye-tracked evaluation [8], interval type evaluation [9], best-selection type evaluation [2] and so on, are stored. Label the set composed by these data as $BD = \{I_U|i=1,2,\cdots\}$, where $I_U$ is the preference information about the user $U_i$. Therefore, by taking the above dataset as a pool of users’ preference, valuable knowledge can be obtained and then be made use to accelerate the algorithm and to improve the optimum quality.

The scheme of BD_IEC is shown in Fig. 2. In order to decrease $\Delta I_{M,U}(t)$, the information difference between the user and the system, the accumulated users’ preference, which composed of the big data, are utilized. First, according to the information $I_M(t,U)$, the user’s preference model $M_{U_i}(t)$ is established. After that, the model feature $\Theta_U(t) = \Theta(M_{U_i}(t))$ is abstracted. Secondly, compare $\Theta_U(t)$ with the sorted elements $\Theta_{U_i}$ in preference features repository. Thirdly, features matching algorithms are used to get the optimum of $\arg \min d(\Theta_U(t), \Theta_{U_i})$. Fourthly, transfer the corresponding information to compensate the user’s preference or to estimate the necessary absent information. Fifthly, the system actively select solutions, which could maximally eliminate the uncertainty for the system, to participate in the evolutionary process to get the user’s preference.

By BD_IEC, the information that evolution system can be used is not only $I_M(t,U)$, but also $BD = \{I_U\}$.

Figure 2. Scheme of big data supported interactive evolutionary computation.

The proposal BD_IEC is based on the hypotheses that under the same condition, the quality of the optimum attained is proportional to both the amount and the quality of dataset about users’ preference.

It is easy to understand that problems with similar features also have similar optimum with a high probability. Those previously conquered problems with their optima stored in the big data can be
made use to accelerate the evolution and to increase the optimum quality for the current optimization problem. The optima stored in the big data can be calculated according to:

$$\{x_T\} = \arg \min \ d(\Theta_U(t), \Theta_{U_i}).$$  \hspace{1cm} (2)

where $x_T$ is the corresponding optimum to the problem feature which is nearest to that of the current optimized problem and is to be transferred to the current optimization. This is feasible for the reason that the optima of those problems with similar features also has similar optimum.

With transfer learning aided BD_IEC, the system can make use both the current user’s preference dataset and the other similar users’ preference data.

Denote the users whose preference are stored in big data repository as $U_{BD} = \{u_i | I_{M_*U_i} \in BD\}$, population in $l-th$ generation as $P(l)$, the evolutionary operators as $E(\cdot)$, then there is:

$$P(t+1) = E(I_M(t, U)) \cup T(\{(x_o, f_{u_i}(x_o)) | u_i \in U_{BD}\})$$ \hspace{1cm} (3)

This means that the $(t+1)-th$ population is generated based on both evolutionary operators and TL aided BD.

In IEC, it is well known that the amount of the user’s preference information is limited by users’ fatigue, therefore, the quality of the information about the search space is vital. If the information is noisy[10], then the model of the users’ preference is not accurate.

In IEC, the information that the system attains is not determined by the information quality but by the promising fitness, i.e. the solutions in new population proposed to user is determined by evolutionary operators, which obey the survival of fittest and other evolutionary rules.

But in a long run, those solutions which can eliminate the uncertainty of information that system attains about the user’s preference is not always less promising than those solutions that evolutionary operators generate. This is true especially under the condition that, the information of users’ preference is limited by users’ fatigue. Therefore, the higher quality the information about users’ preference, the better the algorithm performs. This is also the base for active learning based BD_IEC. Here the quality of the information is measured by the uncertainty can be most eliminated.

Label the certainty of the system about the user’s preference in generation $t$ as $C(t)$, the improvement of the certainty that an individual $x$ in in generation $t$ will bring is $\Delta C(t, x)$. Then active learning will select the individual:

$$\{x_A\} = \arg \max_{x \in \{s \setminus \bigcup_{i=1}^{P(t)} P(i)\}} \Delta C(t, x)$$ \hspace{1cm} (4)

where $s \setminus \bigcup_{i=1}^{P(t)} P(i)$ is the search space as candidate individual, which is expected to accelerate the algorithm by decreasing the uncertainty. In implementation, the uncertainty is generally measured by entropy, i.e. the more entropy, the more uncertainty is. An actively selected solution will decrease the entropy for the user’s preference in the system.

For active learning, those history solutions that the user has evaluated compor of labeled dataset and other solutions in the search space belongs to the unlabeled dataset. Those solutions, that are important for the system to determine which model in repository is more close to the one of the current user, are filtered out. Those actively select solutions are located in search space and are embedded in the new population for the user’s evaluation. With AL aided BD_IEC, there is:

$$P(t+1) = E(I_M(t, U)) \cup A(I_M(t, U), \{I_{U_i}\})$$ \hspace{1cm} (5)

where $I_{U_i}$ is the information that stored in big data as the set of users’ preference. This means that the $(t+1)-th$ population is generated based on both evolutionary operators and TL aided BD.

**Scheme of Big Data Supported IEC**

The scheme of BD_IEC includes includes transfer learning and active learning. Transfer learning carries out the optimum transfer to accelerate the optimization and active learning carries out
solutions recommendation for establishment of the user’s model by which the system can generate promising solutions and improve the users’ satisfaction for the user.

For transfer learning and active learning equipped BD_IEC, the solutions in next generation are from the results of the evolutionary operators, the transferred optimum and the actively selected individuals by active learning, therefore, there is:

\[
P(t + 1) = E(I_M(t,U) \cup \{x_T\} \cup \{x_A\})
\]  

(6)

Experiments

We carried out experiments by developing system based on jMetal[24], which is an object-oriented Java-based framework for single and multi-objective optimization and a number of classic and modern optimization algorithms were implemented.

The design of web page is a time consuming, expensive and complex process. Web page is generally coded with HTML. The presentation semantics of HTML is generally defined by Cascading Style Sheets (CSS), which defines the elements' properties, such as positioning, layout, color and fonts[12]. Therefore, the optimization of HTML page style is based on CSS.

The purpose of the page optimization is to get ones' satisfactory style. Different users have different preference for the style of the same web page. According to a user's subjectively satisfactory style, his/her preferred property values in CSS can be figure out and the values can be further used in other pages.

In our experiments, we only consider the case that different users are faced with the same search space and to optimize the same web page for different style. The task is to optimize the values of properties in CSS. For the users who participated in the optimization, their preference as the evolution history information is stored in the big data repository.

We code CSS value, which determine the style of the pages (or part of the pages), as genotype in genetic algorithm. Figure 3 illustrates the mechanism.

![Figure 3. Mechanism of the optimization for web pages.](image)

The parameters of the experiments are as follows. Two algorithms being compared are: traditional Interactive Genetic Algorithm (IGA) and BD_IQA, and the population size for both algorithms is 20. The crossover operator for both algorithms is simulated-binary-crossover [13] with probability of 0.9 and the mutation operator for both algorithms is polynomial mutation [14] with probability of 0.1. The selection operator for both algorithms is the binary tournament selection [15]. The algorithm termination criterion for both algorithms is: the satisfactory solution is attained. The length of the string that encodes the property value of CSS is 8, and the search space is 256. This is the case that the problem is not large scale and optimized with big data based technique. The experiments were independently carried by 42 students.
The generation numbers for users to get the satisfactory design style are shorten by BD_IGA. Table 1 gives the comparison of the generation number of the two algorithms. One can see that BD_IGA performs better that traditional IGA. What more, from the subjective evaluation from user shows that by BD_IGA, they get more satisfactory solutions than traditional IGA. The subjective evaluation for the solutions are also better than that in traditional IGA.

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional IGA</td>
<td>5</td>
<td>25</td>
<td>19.3</td>
<td>3.54</td>
</tr>
<tr>
<td>BD_IGA</td>
<td>3</td>
<td>14</td>
<td>8.34</td>
<td>2.38</td>
</tr>
</tbody>
</table>

**Conclusion**

We take the many-years accumulated dataset of the evaluation in IEC as the pool of information about users’ preference, and the big-data supported IEC, which is equipped with TL and AL were studied. The simulated experiments validate the efficiency and it shows that BD_EC and BD_IEC are more related with volume of the model in big data, but less related with problems’ difficulty. When the similar or the same problem is collected in big data before the optimization, the performance of the algorithm will be highly improved.

**Acknowledgement**

This research was financially supported by the National Science Foundation of China under Grants No. 61673196, 61503165, 61305149, 61403174.

**References**


