ABSTRACT
Given the importance of optimization and informatics which are the two broad fields of research, we present an instance of Optinformatics which denotes the specialization of informatics for the processing of data generated in optimization so as to extract possibly implicit and potentially useful information and knowledge. In particular, evolutionary computation does not have to be entirely a black-box approach that generates only the global optimal or good quality solutions. How the solutions are obtained in evolutionary search may be brought to light through Optinformatics. In this paper, we present a Frequent Schemas Analysis (FSA) technique for extracting knowledge from the search process by using the historical optimization data, which are otherwise often discarded. FSA brings about greater understanding of GA dynamics through mining for frequent schemas that exists implicitly within the optimization data via the design of frequent pattern techniques (LoFIA) in informatics. To illustrate the principle of optinformatics, a case study using the Royal Road problem is used to explain the search performance of Genetic Algorithm (GA) in action.

Categories and Subject Descriptors: I.2.m [Artificial Intelligence]: Miscellaneous

General Terms: Algorithms, Performance, Experimentation

Keywords: Genetic Algorithms, Frequent Pattern Mining, Schema Theory, Royal Road problem

1. DEFINITION OF FREQUENT SCHEMA
Let function \( Freq(s, t) \) define the frequency of the schema \( s \) in the population at generation \( t \) and \( Freq(s, [m, n]) \) denote the frequency of schema \( s \) in the populations over generations \( m \) to \( n \).

\[
Freq(s, [m, n]) = \frac{\sum_{t=m}^{n} Freq(s, t)}{(n - m)}
\]  

(1)

We define a schema \( s \) as frequent schema with a level \( \theta \) in the period \( [m, n] \) if and only if \( Freq(s, [m, n]) \geq \theta \). One possible interpretation of a frequent schema \( s \) is that GA has spent at least \( \theta \) percentage of its sampling budget on the hyperplane defined by \( s \); or \( \theta \) is a lower bound of the probability that a point in the hyperplane \( s \) is sampled by GA during the period \( [m, n] \).

As stated in Holland’s book [1], "... if some schema begins to occupy a large fraction of the population (through consistent above-average performance), its rate of increase will come very close to \([\mu_{2}(t)/\mu(t)] - 1^\circ\), it is expected that the frequencies \( Freq(s, t) \) of a schema with consistent above-average performance in a period will form a non-decreasing sequence and the set of consistently above average schemas would likely contribute to the set of frequent schemas.

2. FREQUENT SCHEMAS ANALYSIS

In our technique of frequent schemas analysis (FSA) as shown in Figure 1, data which is collected from the evolution process is divided into consecutive and non-overlapping periods. The sampling of GA in each period of the search space is analyzed by investigating the set of frequent schemas \( (Freq(s, P) \geq \theta) \) found in that period. Alternatively, frequent schemas can also be compared across periods to understand the change in GA dynamics. Large value of \( \theta \) gives more confidence on the located convergence regions but the frequent schemas are generally less specific (lower order schemas), thus, interesting information may be not captured.

Each chromosome (binary string) in the data generated by GA in the period is first transformed to a set of items, so as to allow a two-way transformation from a chromosome or schema to an item-set and vice versa. From the possibly numerous frequent schemas, it is up to the analyzer to select interesting schemas from the pool to investigate. In this paper, the interestingness metric is defined as the longest frequent schema (LFS) which provides a sketch on how GA progressively reduces the number of dimensions of its search space or biases its search towards promising regions. Most specific frequent schemas are then found using our LoFIA algorithm which employs bottom-up and depth-first approach to quickly identify the longest frequent schemas from the optimization data. A visualization method is also introduced to capture the change of the schemas across the periods of evolution. Scalar vector \( x \) of length \( L \) represents the set of \( M \) most specific frequent schemas. The value of element \( x_{i} \) for loci \( i \) is then calculated by \( x_{i} = \frac{N_{i}}{N_{0}} \) where \( N_{i} \) and \( N_{0} \) are the number of schemas in the set has value 1 and 0, respectively, at loci \( i \). Vectors \( x \) of consecutive periods are plotted.
3. FREQUENT SCHEMA ANALYSIS OF GA ON ROYAL ROAD PROBLEM

It is worth noting that Random Mutation Hill Climbing (RMHC) outperforms GA on the Royal Road problem. Table 1 shows the average number of evaluations incurred by each algorithm in reaching the optimal solution on the problem of 32 bits (block size $K = 4$) and 64 bits (block size $K = 8$) over 50 independent runs. Our configuration of GA is one-point crossover $p_{cross} = 0.8$, bit-flip mutation $p_{mut} = 0.003$ and fitness-proportional selection with $popsize = 50$.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Royal Road (32K)</th>
<th>Royal Road (64K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>7587.32 ± 7045.26</td>
<td>102880.96 ± 71723.45</td>
</tr>
<tr>
<td>RMHC</td>
<td>412.22 ± 206.61</td>
<td>5876.86 ± 2595.55</td>
</tr>
</tbody>
</table>

Table 1: Hill-climbing outperforms GA on Royal Road problem

against the time axis in the final visualization. Through this visualization, the plot of one period displays the current convergence regions of GA and the differences observed across periods serves to provide hints to the dynamics of GA.

4. CONCLUSIONS

In this paper, a Frequent Schemas Analysis (FSA) technique, which takes its roots from informatics, is introduced for analyzing GA dynamics through mining of frequent schemas that exists implicitly within archived optimization data. In particular, FSA is used to mine for interesting frequent schemas from Binary GA data that is often discarded and investigate the schemas of different search periods visually. FSA provides a comprehensive picture of how the search process evolves, hence bringing new insights into the properties of GA on different problem landscapes. Using the Royal Road problem, we demonstrated the ability of FSA in identifying the premature convergence of GA search which is also confirmed by previous studies. Note that FSA represents an instance of Optiminformatics which aspires to make the evolutionary search more transparent instead of being an entirely black-box approach that serve only to provide good quality solutions.

5. REFERENCES