The Distinguishing String Selection Problem

Anna Gorbenko

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
gorbenko.ann@gmail.com

Vladimir Popov

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
Vladimir.Popov@usu.ru

Abstract

In this paper we consider an approach to solve the distinguishing string selection problem. This approach is based on constructing logical models for the problem.

Keywords: the distinguishing string selection problem, logical model, satisfiability problem, NP-complete

Different sequences algorithms are extensively used in bioinformatics (see e.g. [1] – [3]). In this paper we consider the distinguishing string selection problem (DSS). It should be noted that this problem has important applications in bioinformatics [4]. Also, DSS can be used in some other areas (see e.g. [5]). Note that both the Farthest String and Closest String Problems are special cases of DSS. Therefore, DSS is NP-complete for strings over any alphabet \(\Sigma\) with \(|\Sigma| \geq 2\) [6].

Encoding NP-hard problems as instances of SAT (see e.g. [7] – [11]) and solving them with efficient SAT solvers (see e.g. [12], [13]) has caused considerable interest. In this paper, we consider an approach to solve DSS. This approach is based on constructing logical models for the problem.

Let \(\delta(X,S)\) be the Hamming distance. The decision version of DSS can be formulated as following.

**The distinguishing string selection problem (DSS):**
INSTANCE: Given two sets $\mathcal{M}_c$ and $\mathcal{M}_f$ of strings of length $n$ over an alphabet $\Sigma$, two positive integers $C$ and $F$.

QUESTION: Is there a string $X$ of length $n$ over $\Sigma$ such that $\delta(X, S) \geq F$ for any $S$ in $\mathcal{M}_f$ and $\delta(X, T) \leq C$ for any $T$ in $\mathcal{M}_c$?

Let $\mathcal{M}_f = \{S_1, S_2, \ldots, S_p\}$ and $\mathcal{M}_c = \{T_1, T_2, \ldots, T_q\}$. We use $S[i]$ to denote the $i$th letter in string $S$. Let $\Sigma = \{a_1, a_2, \ldots, a_m\}$. Let

$$\varphi[1] = \land_{1 \leq i \leq p} \land_{1 \leq j \leq n} \lor_{1 \leq k \leq m} x[i, j, k],$$

$$\varphi[2] = \land_{1 \leq i \leq p} \land_{1 \leq j \leq n} \land_{1 \leq k[1] < k[2] \leq m} (\neg x[i, j, k[1]] \lor \neg x[i, j, k[2]]),$$

$$\varphi[3] = \land_{1 \leq i \leq p} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m, s[j]=a_k} x[i, j, k],$$

$$\varphi[4] = \land_{1 \leq i \leq p} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m, s[j] \neq a_k} \neg x[i, j, k],$$

$$\varphi[5] = \land_{1 \leq i \leq q} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m} y[i, j, k],$$

$$\varphi[6] = \land_{1 \leq i \leq q} \land_{1 \leq j \leq n} \land_{1 \leq k[1] < k[2] \leq m} (\neg y[i, j, k[1]] \lor \neg y[i, j, k[2]]),$$

$$\varphi[7] = \land_{1 \leq i \leq q} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m, s[j]=a_k} y[i, j, k],$$

$$\varphi[8] = \land_{1 \leq i \leq q} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m, s[j] \neq a_k} \neg y[i, j, k],$$

$$\psi[1] = \land_{1 \leq j \leq n} \land_{1 \leq k[1] < k[2] \leq m} (\neg w[j, k[1]] \lor \neg w[j, k[2]]),$$

$$\rho[1] = \land_{1 \leq i \leq p} \land_{1 \leq s \leq F} \lor_{1 \leq j \leq n} u[i, s, j],$$

$$\rho[2] = \land_{1 \leq i \leq p} \land_{1 \leq s \leq F} \land_{1 \leq j[1] < j[2] \leq n} (\neg u[i, s, j[1]] \lor \neg u[i, s, j[2]]),$$

$$\rho[3] = \land_{1 \leq i \leq p} \land_{1 \leq s \leq F} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m} (\neg u[i, s, j] \lor \neg x[i, j, k] \lor \neg w[j, k]),$$

$$\rho[4] = \land_{1 \leq i \leq q} \land_{1 \leq s \leq C} \lor_{1 \leq j \leq n} v[i, s, j],$$

$$\rho[5] = \land_{1 \leq i \leq q} \land_{1 \leq s \leq C} \land_{1 \leq j[1] < j[2] \leq n} (\neg v[i, s, j[1]] \lor \neg v[i, s, j[2]]),$$

$$\rho[6] = \land_{1 \leq i \leq q} \land_{1 \leq s \leq C} \land_{1 \leq j \leq n} \land_{1 \leq k \leq m} ((\neg v[i, s, j] \lor \neg y[i, j, k]) \land (\neg v[i, s, j] \lor \neg w[j, k])),$$

$$\xi = (\land_{i=1}^{8} \varphi[i]) \land (\land_{j=1}^{2} \psi[j]) \land (\land_{k=1}^{6} \rho[k]).$$

It is clear that $\xi$ is a CNF. It is easy to check that $\xi$ gives us an explicit reduction from DSS to SAT. By direct verification we can check that

$$\alpha \Leftrightarrow (\alpha \lor \beta_1 \lor \beta_2) \land (\alpha \lor \neg \beta_1 \lor \beta_2) \land (\alpha \lor \beta_1 \lor \neg \beta_2) \land (\alpha \lor \neg \beta_1 \lor \neg \beta_2),$$

$$\land_{j=1}^{l} \alpha_j \Leftrightarrow (\alpha_1 \lor \alpha_2 \lor \beta_1) \land (\land_{i=1}^{l-4} (\neg \beta_i \lor \alpha_{i+2} \lor \beta_{i+1}) \land (\neg \beta_{l-3} \lor \alpha_{l-1} \lor \alpha_l),$$

$$\land_{j=1}^{l} \beta_j \Leftrightarrow (\beta_1 \lor \beta_2 \lor \alpha_1) \land (\land_{i=1}^{l-4} (\neg \alpha_i \lor \beta_{i+2} \lor \alpha_{i+1}) \land (\neg \alpha_{l-3} \lor \beta_{l-1} \lor \beta_l).$$
Table 1: Experimental results for reduction to 3SAT where OA1 (NT) is OA from [12] for natural instances, OA1 (AT) is OA from [12] for algebraic instances, OA2 (NT) is OA from [13] for natural instances, OA2 (AT) is OA from [13] for algebraic instances.

<table>
<thead>
<tr>
<th></th>
<th>OA1 (NT)</th>
<th>OA1 (AT)</th>
<th>OA2 (NT)</th>
<th>OA2 (AT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>49.7 min</td>
<td>17.3 min</td>
<td>12.4 min</td>
<td>2.6 h</td>
</tr>
<tr>
<td>maximum</td>
<td>7.3 h</td>
<td>5.8 h</td>
<td>23.8 min</td>
<td>18.9 h</td>
</tr>
<tr>
<td>best</td>
<td>12.2 min</td>
<td>9.1 min</td>
<td>51 sec</td>
<td>16.6 min</td>
</tr>
</tbody>
</table>

\[ \alpha_1 \lor \alpha_2 \Leftrightarrow (\alpha_1 \lor \alpha_2 \lor \beta) \land \\
(\alpha_1 \lor \alpha_2 \lor \neg \beta), \quad (3) \]

\[ \bigvee_{j=1}^{l} \alpha_j \Leftrightarrow (\alpha_1 \lor \alpha_2 \lor \beta_1) \land \\
(\neg \beta_1 \lor \alpha_3 \lor \alpha_4) \quad (4) \]

where \( l > 4 \). Note that using relations (1) – (4) we can easily obtain an explicit transformation \( \xi \) into \( \zeta \) such that \( \xi \Leftrightarrow \zeta \) and \( \zeta \) is a 3-CNF. It is easy to see that \( \zeta \) gives us an explicit reduction from DSS to 3SAT.

In this paper we consider 3SAT solvers from [12], [13]. We have created a generator of natural instances for DSS. Also, we have created a generator of algebraic instances for DSS. Algorithmic problems of algebra have been studied intensively in the last decades (see e.g. [14] – [20]). In particular, avoidance of words plays an important role in solving such problems (see e.g. [21] – [32]). It is easy to see that DSS can be used for generation of avoidable words. Our generator of algebraic instances creates instances for generation of avoidable words. We have used heterogeneous cluster for our computational experiments. Each test was runned on a cluster of at least 100 nodes. Selected experimental results are given in Table 1.

References


Received: May, 2012