Classifying with Decision Diagrams
Obtained from Genetic Programming

Robin Nunkesser
Faculty of Computer Science, TU Dortmund
1 Introduction
   • Motivation
   • Evolutionary Computation
   • Problem Definition

2 The Algorithm
   • Definition

3 Results
   • Genetic Association Studies
Genetic Association Studies

**Task:** Identify genetic factors that may contribute to a medical condition

- Data sets on genetic factors of individuals (*cases* and *controls*)
- SNPs are the genetic factors typically considered

**Single Nucleotide Polymorphism**

DNA sequence variation

- **Single:** located at a single base/nucleotide pair
- **Nucleotide:** sugar, phosphate and base
  - Adenine, Thymine,
  - Cytosine, or Guanine
- **Polymorphism:** nucleotide variant with a frequency of $\geq 1\%$

$\approx 90\%$ of genetic variations are SNPs
Single Nucleotide Polymorphisms

- Most SNPs are biallelic (two variants exist)
- Let $A$ be the major/reference allele and $a$ the minor allele
- Biallelic SNPs can be divided into three types determined by the mother chromosome and the father chromosome:
  - homozygous reference $AA$ (coded as 0)
  - heterozygous variant $aA$ or $Aa$ (1)
  - homozygous variant $aa$ (2)

Data Example

<table>
<thead>
<tr>
<th>SNP₁</th>
<th>SNP₂</th>
<th>SNP₃</th>
<th>SNP₄</th>
<th>SNP₅</th>
<th>SNP₆</th>
<th>case/control</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Evolutionary Computation (EC)

- General term for the usage of search heuristics inspired by natural evolution
  - possible solutions are represented by *individuals* (describing elements in a *search space*)
  - a set of individuals (a *population*) undergoes variation (*crossover* and *mutation*)
  - the *fitness* of the individuals is evaluated
  - a new generation is derived after a *selection* process

- Black-box optimization

- Often used for problems that are not easy to solve by conventional methods
  - Combinatorial optimization
  - Learning
  - Genetic association studies, robust regression, evolutionary clustering, time series modeling...
1. Create an initial random population.
2. Evaluate the fitness values of the population.
3. Perform the following steps on the current generation:
   1. Select individuals in the population based on a selection scheme.
   2. Adapt the selected individuals.
   3. Evaluate the fitness value of the adapted individuals.
   4. Select adapted individuals for the next generation according to a selection scheme.
4. If the termination criterion is fulfilled, then output the final population. Otherwise, set the next generation as current and go to step 3.

Graphical Representation
Problem Definition

Output in $B := \{0, 1\}$, inputs from $P := \{0, \ldots, p\}$

**Goal:** Find the function $f : P^n \rightarrow B$ that “fits” the input best

**Our approach:** Map multiple-valued variables to all distinct Boolean variables

\[
x^a := \begin{cases} 
1, & \text{if } x = a \\
0, & \text{otherwise}
\end{cases}
\]

and $\overline{x}^a$ and search for polynomials

---

**Example of an individual**

![Decision Diagram Example]
New Problem

Output in $Q := \{0, \ldots, q\}$, inputs from $P := \{0, \ldots, p\}$

Multiterminal Multivalued Decision Diagrams

- Start evaluation at the root
- Follow the corresponding edges
- Sinks determine the class

Classify two ternary values
Genetic Programming Algorithm

1. Create an initial random population

2. Perform the following steps on the current generation:
   1. Select all individuals in the population, and reproduce them
   2. Conduct the following adaptations on randomly selected individuals
      1. Change the variable of a node
      2. Add a new node to the decision diagram
      3. Remove a random node
      4. Redirect a random edge
      5. Conduct a special mutation

3. Evaluate the fitness values as
   1. Number of correctly classified observations
   2. Size of the decision diagram

4. Remove pareto dominated individuals

If the termination criterion is fulfilled, then output the final population. Otherwise, set the next generation as current and go to step 2
Special Mutation

1. Choose a random number of observations
2. Mutate the current classification for these observations
3. This operation is very fast in ordered decision diagrams

**Definition**

A *variable ordering* $\pi$ is a permutation on the index set $\{0, \ldots, n\}$. An *ordered decision diagram* is a decision diagram, where the sequence of tests on a path is restricted by the variable ordering $\pi$. 
Results on Real Data

HapMap
- Original purpose: Search for common haplotypes of four different ethnies
- Here: Discriminate between four ethnies
- 121774 SNPs from 90 individuals

GENICA
- Gene ENvironment Interaction and Breast CAncer in Germany
- Survey on genetic and environmental influence factors on sporadic breast cancer
- Here: 63 SNPs from 1258 women
Results on HapMap

Pareto front approximation

Observations

- Groups are completely separable with \( \approx 20 \) nodes
- Smaller and more interesting diagrams are identified
GENICA

- Is a case/control study
- The original model of polynomials is more natural
- Decision Diagrams from our algorithm deliver similar results

Diagram with 5 nodes

- CYP1B1_1358
- ERCC2_6540
- ERCC2_18880
- control
- case
Simulation (R package scrime)
- Response by level-wise logistic regression models
- Level 0: $\bar{x}_6^0 x_7^0$
- Level 1: $x_3^0 x_9^0 x_{10}^0$
- Level 2: $\bar{x}_2^0 x_5^0 \vee x_1^0 x_4^0 x_8^0$
Constructing Decision Diagrams with Genetic Programming

- allows efficient operations,
- works in the presence of more than two classes,
- provides a nice alternative for the case of two classes.
URL http://doi.acm.org/10.1145/1389095.1389339

URL http://dx.doi.org/10.1093/bioinformatics/btm522