Diagnosis of SIDS using Genetic Programming

Skobtsov Y.A., Vasyaeva T.A.

IAMM, Rose Luxemburg str., 74, Donetsk, 83114, UKRAINE, DNTU, Artema str.,58, Donetsk, 83000, UKRAINE, E-mail: Skobtsov@kita.dgtu.donetsk.ua, vasyaeva_tanya@tr.dn.ua

Abstract – Genetic Programming is a relatively new technique for the automatic discovery of computer programs which offer solutions to complex problems. It is being applied to an ever increasing number of application areas with good results. This report presents a work on a classification problem. The Genetic Programming technique is used to evolve programs which are able to diagnose sudden infant death syndrome in a sample of patients. Real data on the habits and lifestyles of patients are used to train and test the system. The system makes accurate predictions and competes with an other methods of this problem.

Key words – Genetic Programming, mutation, crossover, reproduction, terminal set, function set, fitness measure, full and grow trees

I. Introduction

Sudden infant death syndrome (SIDS) [1] is the sudden death of an infant under one year of age which remains unexplained after a thorough case investigation, including performance of a complete autopsy, examination of the death scene, and review of the clinical history.

In a typical situation parents check on their supposedly sleeping infant to find him or her dead. This is the worse tragedy parents can face, a tragedy which leaves them with a sadness and a feeling of vulnerability that lasts throughout their lives. Since medicine can not tell them why their baby died, they blame themselves and often other innocent people.

II.Genetic Programming

Genetic Programming (GP) [2] uses the Darwinian principle of natural selection to produce solutions to problems.

Using GP, it may be possible to apply computer technology to problems which are too complex to solve by a human programmer using conventional methods in computer science.

In GP, the individuals which take part in evolution are computer programs which pose a solution to a given problem. Computer programs can be represented as trees, and the production of successive generations is achieved by applying genetic operators [3].

Evolution is captured in a repetitive computational process. An initial population is generated randomly as a starting point for the process. Each individual is executed to measure its fitness or ability to solve the problem for which it is intended. A new population is then generated from the previous generation, using the genetic operators. The probability that a particular individual participating in the generation of an individual in the new generation, increases with its fitness. This process of fitness evaluation and evolution is repeatedly applied until a solution is found or a timeout occurs.

III.Source Data

The data used in this project was originally collected at the obstetric-gynecologic and forensic medicine departments DNMU in Donetsk.

The data represent results of the examination 240 women. Among them 120 women have died the children from sudden infant death syndrome (SIDS) and checking group from 120 patients with alive children.

The data contain general information and lifestyle pregnant, diseases during pregnancy and results some analysis.

IV.Terminal Set

Terminal set consists of factor of the risk, which are preparation in the following way:

- place of address (town = 1, village = 0);
- age at a point in labor ≤ 17;
- age at a point in labor ≤ 25;
- age at a point in labor ≤ 30;
- age at a point in labor > 31;
- harmful conditions;
- age at a point in first menstruation ≤ 12;
- age at a point in first menstruation ≤ 14;
- age at a point in first menstruation > 15;
- regular menstruation;
- painful menstruation;
- duration to menstruation (number of days) ≤ 3;
- duration to menstruation (number of days) ≤ 5.5;
- duration to menstruation (number of days) > 6;
- etc.

Each factor represents the true or false value.

V. Function Set

The function set consists of three logical operators: AND, OR and NOT. We have changed operation NOT on operations AND-NOT and OR-NOT, so as from now the function set consists of four logical operators: AND, OR, AND-NOT and OR-NOT,

A solution to the problem will therefore be a logical expression using patient attributes as variables.

VI.Fitness Measure

Genetic Programming generates candidate solutions to the problem being solved. It is necessary to have a method of measuring how effectively a program solves the problem. This is known as the fitness measure. It is important that this measure accurately assesses the performance of a program.

The method was considered: proportion of patients correctly diagnosed.

VII.Algorithm

Genetic Programming requires a representation scheme with which it can express a solution to the problem which it is trying to solve. Solutions to GP problems are expressed as programs or trees. These trees are constructed from functions which are the internal nodes of the trees, and terminals which form the leaves of the trees.

The generalized algorithm GP:

1. Installation parameter GP.
2. The generation to initial population. The population is the number of genetic programs. Each genetic program is the tree, which represent decision. The tree, on initial stage is generated in a random way and consists of functional and terminal nodes.
3. Rate fitness measure for each tree in populations.
4. Performing genetic operators. [3]
5. Check criterion stop. If problem is solved or maximum number of generations has been reached - step 6, otherwise - step 3.
6. Choice of the best decision in the last population.

This algorithm is programmed for realization of the delivered task to use C++ Builder 6.

Table 1 lists the values of the main parameters used in the best runs for this report.

TABLE 1

GP PARAMETERS FOR THE SIDS DIAGNOSIS PROBLEM

Parameter	Value
Population size	200
Maximum depth of individuals in population	10
Generative method	ramped half-and-half
Probability of function node	50%
Probability of terminal node	50%
Crossover probability	99%
Mutation probability	5%
Selection method	roulette

The population size is the number of genetic programs in each generation of the genetic programming process. Typically, a more difficult problem would employ a larger population.

The ramped half and half method of population generation attempts to produce a varied first generation with many different types of trees. Half of the generated population will contain full trees which have all their leaf nodes at the same depth. In these trees, a function node is generated at each level until the deepest permissible level is reached, at which point a terminal node is generated. Example full and grow trees are shown in figure 1. The other half of the population contains grow trees. The shape of grow trees is determined probabilistically using the probability of function and terminal node parameters. When creating each node using the grow method, a decision is made as to which type of node; function or terminal; to generate. In this case, 50% of the nodes will be function nodes, and 50% will be terminal nodes. As with full trees, when the maximum depth is reached, the generation of a terminal node is forced.

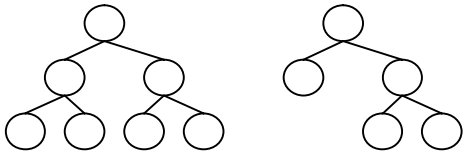

Fig.1 a) Full and b) Grow generated trees

During the generation of each half of the population, the maximum depth of the individual programs is incremented from 4 to the maximum depth (10 in this case), to ensure the widest possible variety of individuals in terms of shape and size. It is worth noting that to further ensure variety in the initial population, a check is performed to ensure that each individual is unique.

They are used consecutively genetic operators [3] reproduction, crossingover, mutation and reduction.

Conclusion

It was necessary to run the GP system several times with different initial populations.

The best solution correctly predicts 95.71%. The part of the best decision are shown in figure 2.

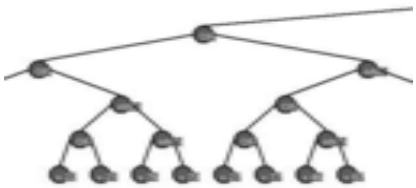

Fig.2 The part of the best tree.

The results have clearly demonstrated that genetic programming can be used effectively for the diagnosis of SIDS from patient questionnaire data.

The results are particularly pleasing in light of the problem of inconsistencies in the source data. This indicates that such a GP system would be of practical use in a real-world diagnosis application where such noise would be inevitable.

Genetic Programming has been shown to effectively diagnose SIDS. There does not appear to be any reason why this should not be extended to the diagnosis of other conditions and diseases in the medical world.

References

[1] http://www.sids.org/index.htm

[2] W. Banzhaf et all. Genetic Programming – an Introduction. – Morgan Kaufman, Heidelberg:San-Francisco, 1998.

[3] Рутковская Д., Пилинский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы: Пер. с польск. И.Д. Рудинского. - М.: Горячая линия – Телеком, 2006. – 452 с. : ил.