Application of Genetic Algorithms in the analysis of Traditional Chinese Medicine data

I-CHENG JOHNY SHIH and JOSIAH POON (Supervisor)
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

INTRODUCTION & MOTIVATION

Herbal medicine is the main treatment method in Traditional Chinese Medicine (TCM). It is prescribed by using a core set of herbs for the main syndrome and then additional herbs to deal with other symptoms that the patient may have. Without using domain knowledge of TCM we aim to extract the core sets and any possible addition sets to reconfirm TCM theories.

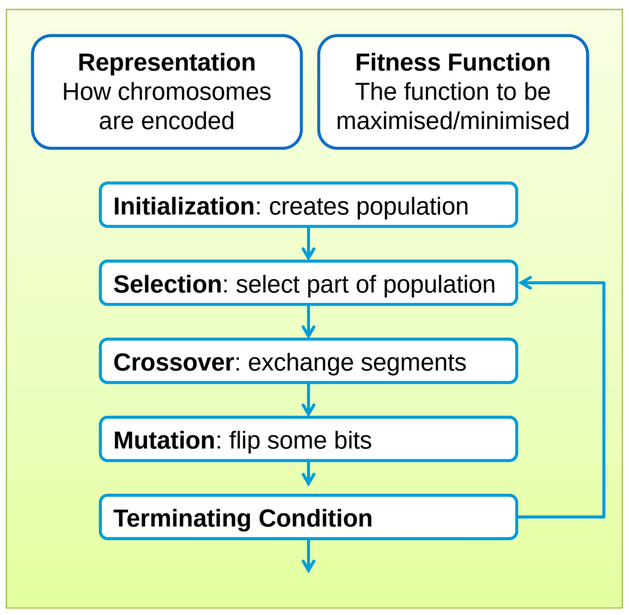
Co-evolutionary Genetic Algorithm (CoGA) allows the task of optimising two populations from different domains as a single population linked by the fitness function. It was successfully applied to find partially significant subsets between symptoms and herbs in a TCM Insomnia dataset [1].

Project Aim

Use GA or variants of GA to find the following in a TCM dataset by creating semantically correct fitness functions.

- Core symptoms and herbs set
- Side effect symptoms and herbs sets

GENETIC ALGORITHM

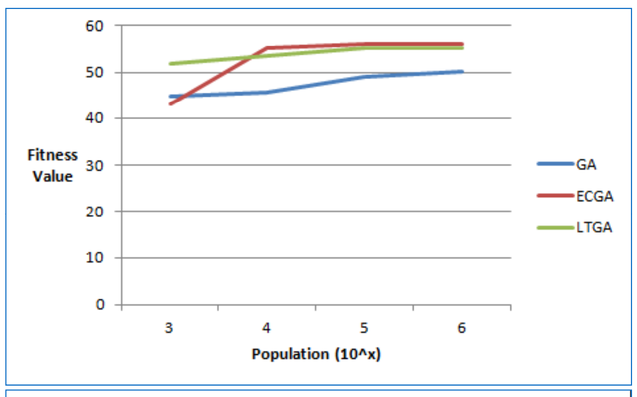


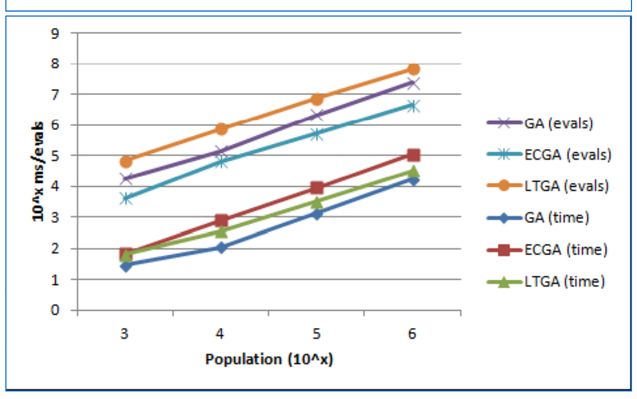
VARIANTS OF GAs

- GA performance can be improved by effective building block mixing (linkage learning); traditional methods such as elitism and niching do not provide significant improvements to scalability [2]. This has led on to the development of a series of linkage learning GAs in hopes of improving the scalability of GA. Two competent variants are used in this project and presented below.
- Extended Compact GA (ECGA) is based on the idea that a good probability distribution is equivalent to linkage learning. Generates solutions using the probability distribution instead of applying standard crossover [3]. ECGA is well researched and frequently used linkage learning genetic algorithm.
- Linkage Tree GA (LTGA) builds an agglomerative hierarchical tree using entropy as the distance measure. This tree is then used in the crossover step by taking each node as a crossover mask. The children are only retained if one of them yields a better fitness than both parents. LTGA's mainly designed to solve additively decomposable problems [4].

COMPARISON OF GAs

GA / ECGA/ LTGA compared using Trap-K fitness function with 6 bit blocks and 8 blocks per chromosome (max score 56).





DATASETS

Insomnia Dataset

- 457 records with binary outcomes.
- 102 symptoms + 111 herbs = 213 attributes.
- Sparse: Average 4 symptoms, 13.5 herbs.
- 85.1% positive outcome

Tourette Syndrome Dataset

- 1391 records with ordinal outcomes.
- 4 syndromes + 189 herbs = 193 attributes.
- Sparse: 1 syndrome, 11 herbs.
- Average Efficacy: 0.341 (scale 0~1).

FITNESS FUNCTIONS

Pairwise Efficacy Gain

$$P(i,j) = 2p(i,j) - (p(i) + p(j))$$

$$F(x) = \frac{\sum_{\forall (i,j) pair \in x, i \neq j} P(i,j)}{\binom{|x|}{2}}$$

where x is the chromosome
i, j are active genes (attributes) within x.
p(i, j) is the average outcome in the dataset when i and j are active

• Measures efficacy gain •Computationally cheap due to |x| being small.	•Requires a min threshold. •Does not take into account number of occurrences.

Pairwise Multi Objective

$$C(i,j) = 2c(i,j) - (c(i) + c(j))$$

$$R(i,j) = \sqrt{\log \frac{c(i,j)}{|n|}}$$

$$P'(i,j) = \frac{P(i,j) + C(i,j)}{2} * R(i,j)$$

$$F(x) = \frac{\sum_{\forall (i,j) pair \in x, i \neq j} P'(i,j)}{\binom{|x|}{2}}$$

where c(i, j) is the number of records in the dataset with i and j being active.
|n| being the number of records in dataset

•Measures efficacy, co-occur chance and penalises low occurrences	•Penalty amount may give too much advantage to frequently occurring genes.

EXPERIMENT & RESULTS

Significant Subsets

Pairwise Efficacy Gain + Insomnia Dataset
Tinnitus, Fever;
Lotus Plumule, Fresh Rehmannia Root, Yam Rhizome, Asiatic Cornelian Cherry Fruit

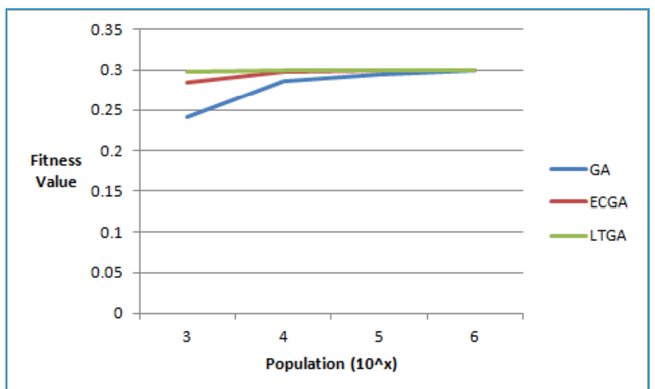
The above set have been identified as a side effect subset by TCM doctors and is in complete accordance with TCM theories. The fitness function's semantics imply that this was the most effective subset from the dataset.

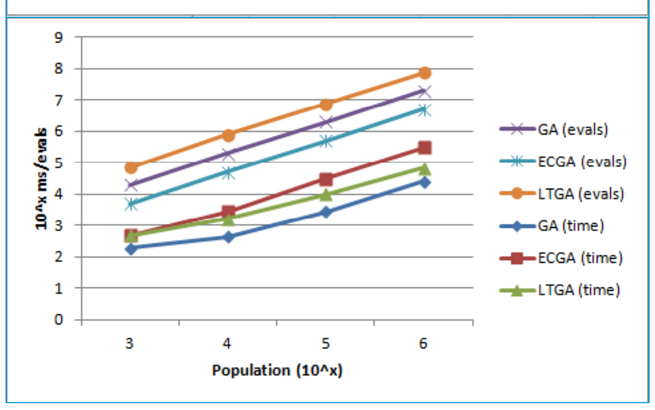
Pairwise Multi Objective + Insomnia Dataset
Difficulty Falling Asleep, Insomnia;
Poria, Licorice, Anemarrhena, Fried Ziziphi.

In contrast to the previous, this was identified as the core set. The herb components form suanzaoren decoction which is commonly used in TCM to treat insomnia.

Performance

The following was done using Tourette dataset with pairwise multi objective.





FUTURE WORK

- Larger datasets are needed to get more reasonable values for significance testing.
- The penalty problem could potentially be solved by using statistical measures inside the fitness functions.

CONCLUSION

Our work shows that GA is a powerful tool for analysing TCM datasets. We found LTGA to perform the fastest given our dataset and fitness function but ECGA holds potential in more complicated fitness functions. With these variants, we were capable to extracting both the core set and some side effect sets of the data that are in accordance to TCM theories.

References

[1] Poon, J; Yin, D; Poon, S; Zhang, R;Liu. B; Sze, D; "Co-evolution of symptom-herb relationship," Evolutionary Computation (CEC), 2012 IEEE Congress on , vol., no., pp.1-6, 10-15 June 2012
[2] Thierens, D; 1999. Scalability problems of simple genetic algorithms. Evol. Comput. 7, 4 (December 1999), 331-352. DOI=10.1162/evco.1999.7.4.331
[3] Thyago S.P.C. Duque; Goldberg, D.E; 2009. A new method for linkage learning in the ECGA. In Proceedings of the 11th Annual conference on Genetic and evolutionary computation (GECCO '09). ACM, New York, NY, USA, 1819-1820. DOI=10.1145/1569901.1570179
[4] Pelikan, M; Hauschild, M.W; Thierens, D; 2011. Pairwise and problem-specific distance metrics in the linkage tree genetic algorithm. In Proceedings of the 13th annual conference on Genetic and evolutionary computation (GECCO '11), Natalio Krasnogor (Ed.).