This project uses the PFGP (Probability-Fix Genetic Programming) algorithm to solve symbolic regression, which is a variant of genetic programming algorithm. The algorithm first calculates the correlation magnitude between features and targets from the original data, then establishes a determined feature selection probability based on the correlation magnitude, and selects new features based on the established feature selection probability during the evolutionary process when a mutation occurs and a new feature needs to be selected. The process of PFGP is shown in Figure 1.

PFGP first calculates the MIC value c_i of each feature and the target value. The MIC value reflects the degree of relevance of the feature to the target, and its value close to 1 indicates that the corresponding feature has a high degree of relevance to the target, then the feature has a higher probability of being selected. the probability p_i of the ith terminal is

$$p_i = \frac{c_i}{\sum_{j=1}^{n} c_j} \tag{1}$$

The obtained feature selection probabilities will remain fixed during the evolution of the PFGP. Only the update of the symbol selection probability is performed during each iteration round. The symbol selection probability is determined based on the number of symbols within the current population, and symbols with more occurrences have a higher probability of being selected in the next evolutionary rounds.

FPGP follows the pattern of traditional genetic programming algorithms for population initialization and evolution. In one evolution, the number of symbolic positions in all chromosomes *FQ* is counted and normalized by the following equation:

$$FQ = \frac{FQ}{POPSIZE * L} \tag{2}$$

where *POPSIZE* is the population size and *L* is the fixed size of each individual. Through equation (2) *FQ* is defined as the percentage of the number of symbols within the population relative to the sum of the number of symbols and the number of features. For each individual, its probability of mutation *CR* and locus *k* are determined randomly. After determining both, a random number is generated first, and if this random number is smaller than *CR* then mutation occurs, otherwise no mutation occurs and the individual is directly retained in the new population. If locus *k* is located at the head of

the chromosome, there is a possibility of conversion between symbols and features, when another random number is generated and compared with *FQ*, if this random number is smaller than *FQ* then a new symbol is selected for replacement by mutation at this locus, and vice versa a feature is selected for replacement. If locus *k* is located in the tail, we determine whether the original feature on locus *k* is a feature, and then select a new feature or symbol based on the feature selection probability or symbol selection probability to form a mutated new individual. By evaluating the fitness of the mutant individual and then comparing it with the R^2 of the father individual, the better of the two is selected to enter the next generation population.

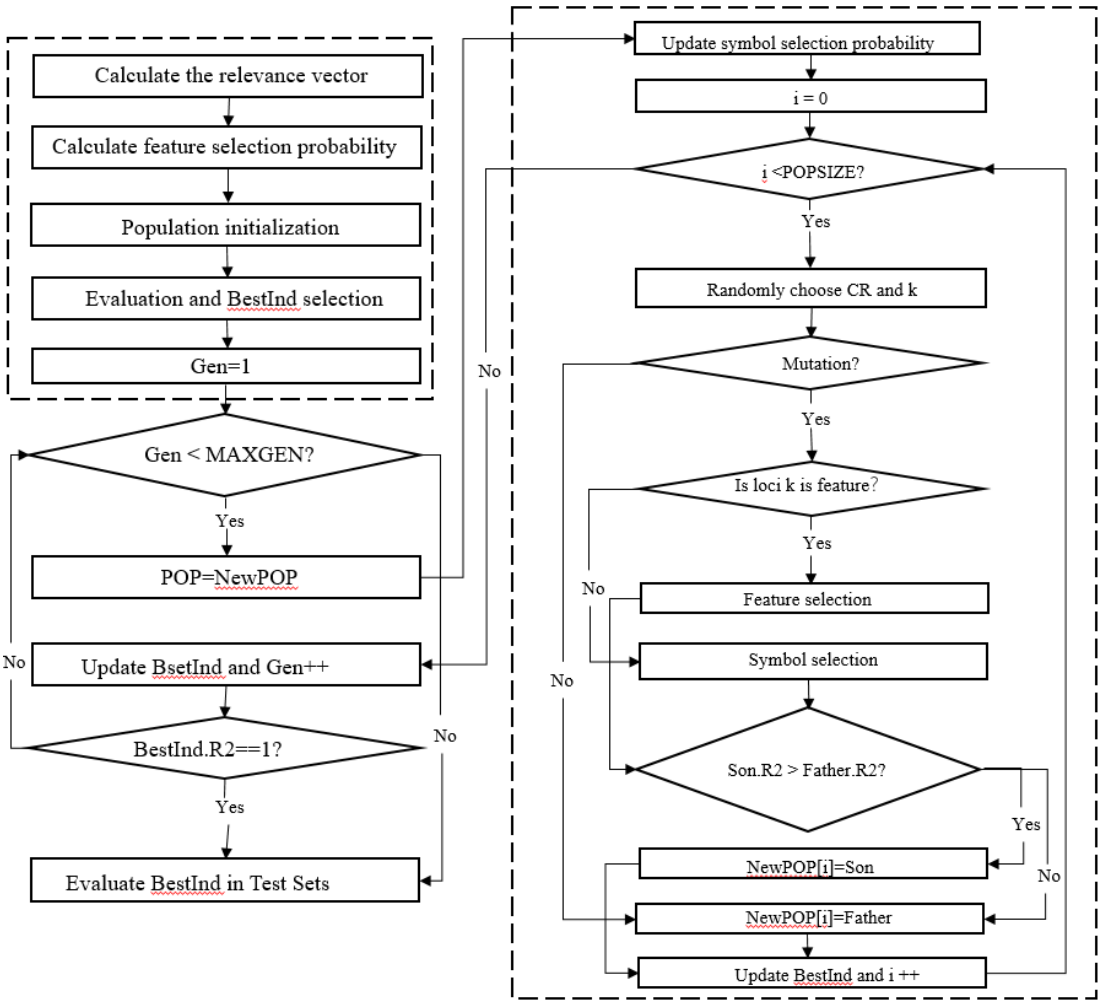

Fig. 1 PFGP pipline

Contact information of the participants：

Group Member 1: Lianjie Zhong, Email： jackyzhong99@qq.com

Group Member 2: Jinghui Zhong, Email: jinghuizhong@suct.edu.cn

Group Member 3: Dongjunlan, Email: eru-dd@foxmail.com

Group Member 4: Nikola Gligorovski, Email:nikola-gligorovski@qq.com