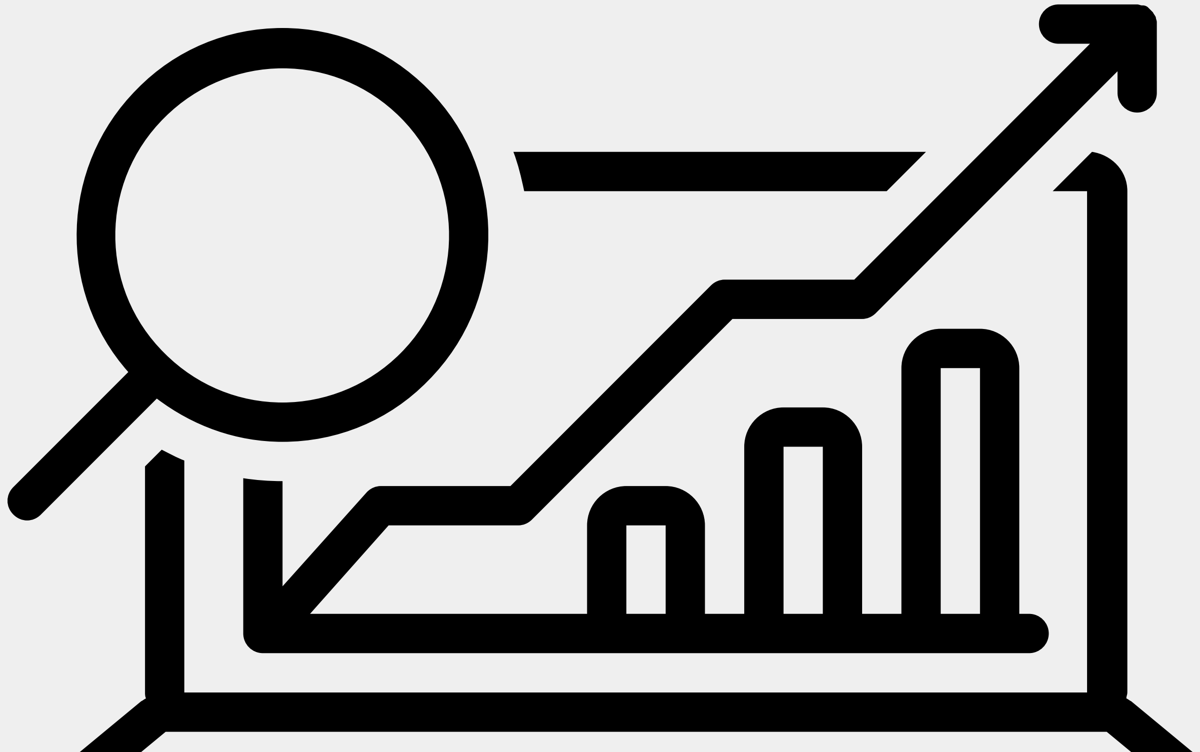
EVOLUTIONARY FEATURE SELECTION APPROACHES FOR INSOLVENCY BUSINESS PREDICTION WITH GENETIC PROGRAMMING

Computational Intelligence Spring 24



Members:
Ali Muhammad Asad
Ali Siddiqui

INTRODUCTION

Introduction

- Feature selection methods in the field on Business Failure Prediction Models
- BFPMs aim to anticipate the difficulties faced by a company
- Provide useful tools for critical decision making
- Insolvency problem considered as a classification problem
- Test the capabilities of Genetic Programming as an appropriate classifier for explanatory variables

THE PROBLEM

The Problem

- Exponential increase in financial information in recent times
- Increased capacity for classification techniques to deal with more variables
- Available datasets can be overwhelmed with a multitude of financial features
- Results in increased time and costs to obtain solutions
- No consensus in BFP to select a variable with the classification method - thus our problem
- Need to discover relevant characteristics

Feature Selection Methods

- **Two methods to select any feature:**
  1. *Filter Approach*: - Does not depend on classifier
     - use measures which only depend on intrinsic data properties
     - features based on statistical measure
  2. *Wrapper Approach:* - depends on specific classifier model
     - train classifier with different character subsets
     - select the best classification performance
- **Filter Approach faster, but ignores intersection of features and performance of characteristics in classification**
- **Wrapper Method poses a high computational burden**

METHODOLOGY / SOLUTION

EA Formulation

Algorithm 1 Differential Evolution algorithm.

Initialize the population of random solutions
Evaluate vectors (encoded solutions) (KNN accuracy with the encoded selected ratios)
repeat
  for all vector x in the population **do**
    Let $x_1, x_2, x_3 \in$ population, randomly obtained $\{x_1, x_2, x_3, x$ different from each other$\}$
    Let $R \in \{1, ..., n\}$, randomly obtained {n is the dimension of the search space}
    for $i = 1$ *to* n **do**
      Pick $r_i \in U(0,1)$ uniformly from the open range (0,1).
      if $(i = R) \vee (r_i < CR)$ **then**
        $y_i \leftarrow x_{1i} + F(x_{2i} - x_{3i})$ {CR - Crossover probability, F - Weight factor}
      else
        $y_i = x_i$
      end if
    end for$\{y = [y_1, y_2 ... y_n]$ is a new generated candidate or trial vector$\}$
    Evaluate fitness $f(y)$ of candidate y (KNN accuracy with its encoded selected ratios)
    if $f(y) \leq f(x)$ **then**
      Replace vector x by y {if y has better or equal fitness}
    end if
  end for
until termination criterion is met
return $z \in$ population $\backslash \forall t \in$ population, $f(z) \leq f(t)$

Fitness Function

The population size is maintained constant by the selection process. The trial vector (y) and the target vector (x) are compared, keeping in the next evolutionary generation the fittest one. In this way, the algorithm incorporates elitism since the best solution (vector) is maintained or improved throughout the generations

Crossover

The result of the crossover operation defines the final
trial vector (y) for each target vector x. The standard “binomial” crossover (specified in Algorithm 1) is used.

GP Parameters

Explanatory variables	The ones corresponding to each selected subset, obtained by the different feature selection methods
Variable transformation	Normalization based on maximum and minimum values of financial ratios
Evaluator	Mean Squared Error (MSE of predicted values with respect to correct values in the training set)
Solution creator	Probabilistic Tree Creator
Symbolic expression tree grammar	Arithmetic functions (+, -, *, /)
Maximum depth	10 (maximum depth of the tree)
Maximum length	100 (maximum length of the symbolic classification model)
Population size	1,500
Maximum generations	100
Crossover	Subtree Swapping Crossover (crossover of subtrees at the crossover point)
Mutation	Multi Symbolic Expression Tree Manipulator (allows different types of mutation)
Mutation probability	15%
Selector	Tournament - Window size 8 (used in mutation and crossover)
Elites	1 (only best solution retained
Model creator	Accuracy Maximizing Thresholds (the returned solution is the one that uses as classification threshold the one that maximizes the percentage of successes in the training set)

RESULTS

DE/KNN Setup

Test Variants:

i. T1-30FS: DE selects from 30 relevant ratios based on Fisher Score

ii. T2-30TS: DE selects from 30 relevant ratios based on T-statistic

iii. T3-59 Ratios: selects from entire set of 59 ratios

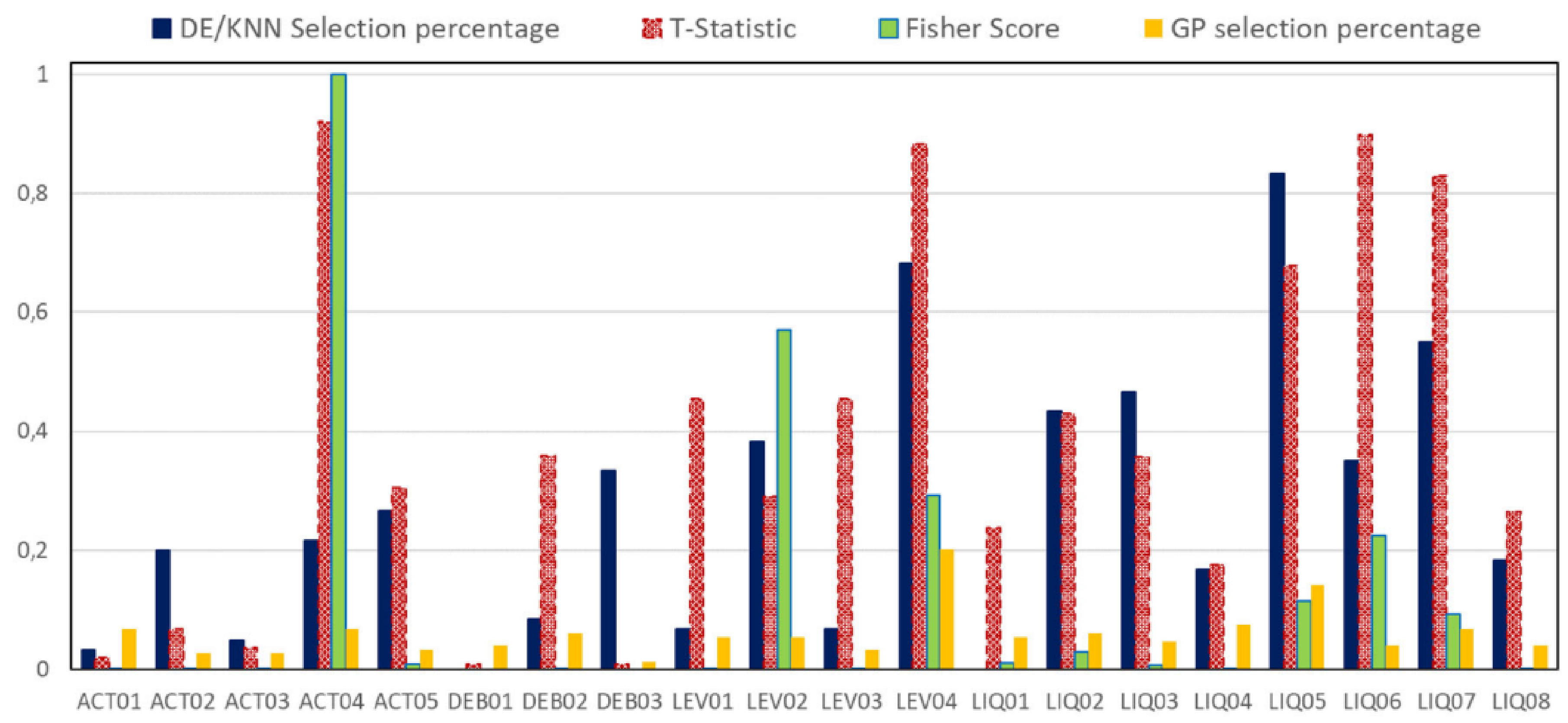
Parameters:

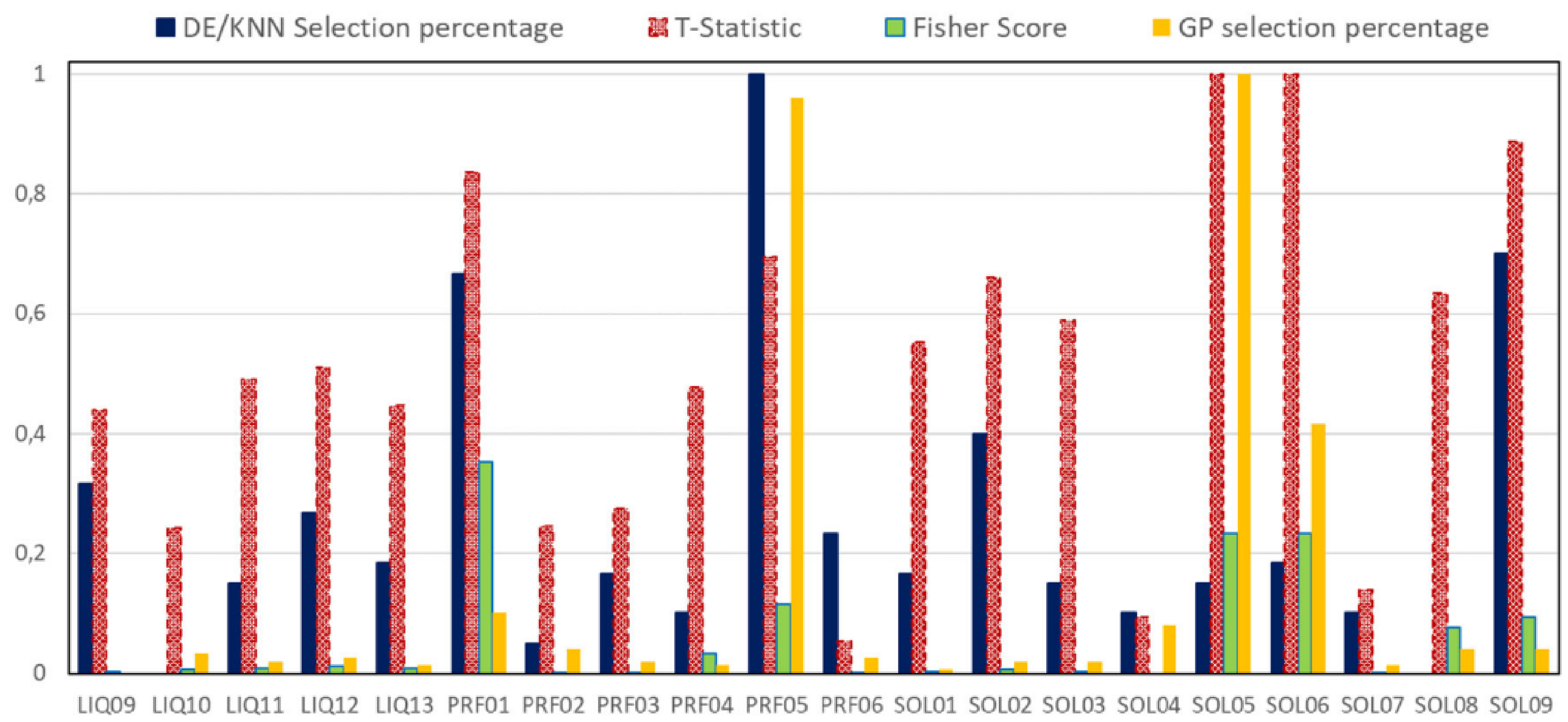
- Population Size = 100
- Low Crossover Probability (0.1)
- F param takes a random value between [0, 9]
- Generations = 500

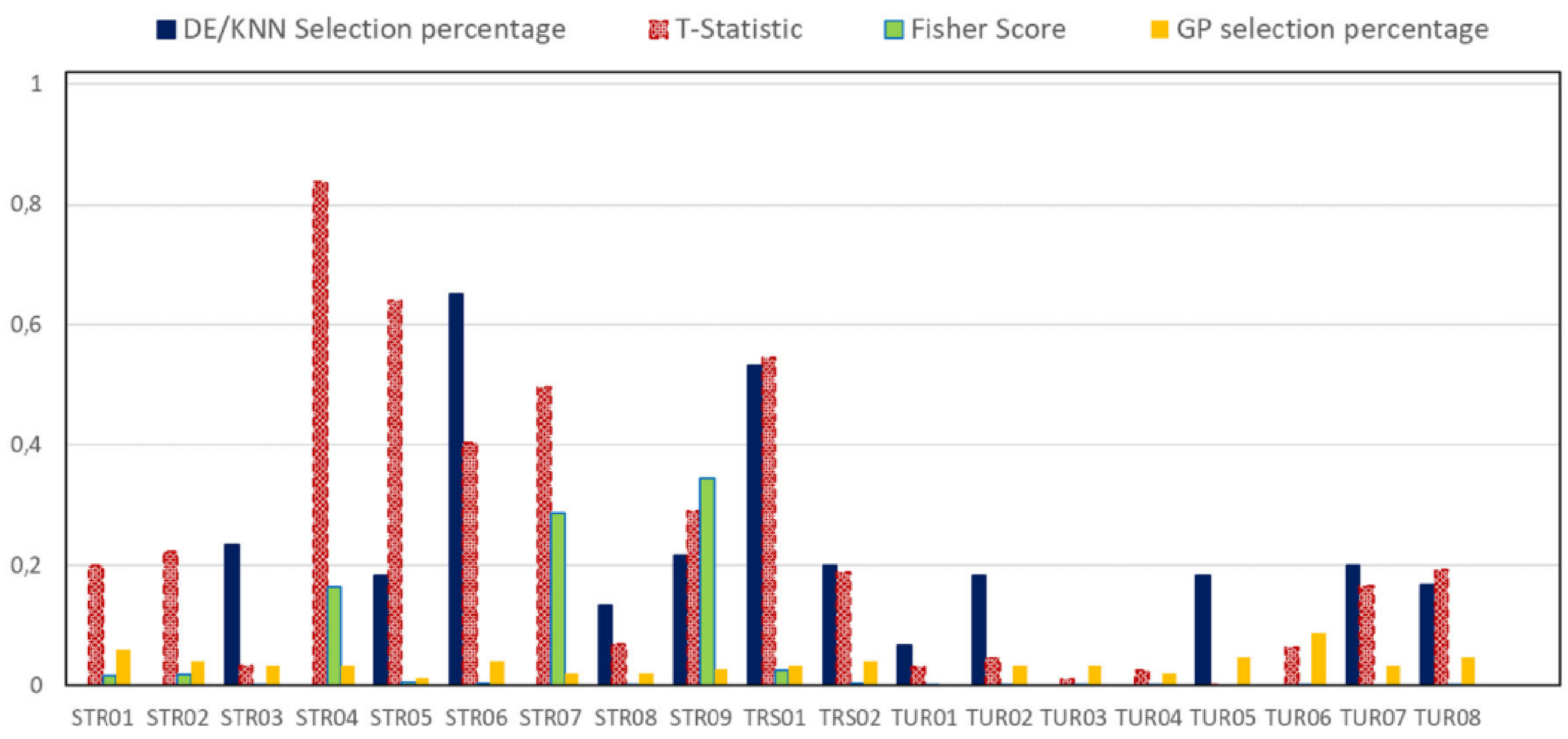
DE/KNN Results

Test variant (starting pool of ratios)	Number de neighbors (KNN)			
	3NN		15NN	
	Average	Best	Average	Best
59 ratios (T3-59 Ratios)	88.48	90.77	88.17	89.63
30 best ratios with Fisher Score (T1-30FS)	86.23	88.6	88.40	89.70
30 best ratios with T-statistic (T2-30TS)	85.87	87.20	87.69	89.44

Table1: Classification accuracy (fitness) in the different test variants







GP Setup

- **Variables:** corresponding to each related subset obtained by varying selection methods (normalized variables)
- **Population Size:** 1500
- **Maximum Generations:** 100
- **Crossover:** Subtree Swapping Crossover (crossover of subtrees at the crossover point)
- **Mutation:** Multi Symbolic Expression Tree Manipulator (allows different types of mutation) using 15% mutation rate
- **Tournament Selection Method**

DE / KNN vs GP

Classification measures using the complete test set, with selected inputs from the DE / KNN Approach

Test variant	Accuracy	Sensitivity
T1-30FS (30 best ratios with Fisher Score), 3NN	87.45	80.88
T1-30FS (30 best ratios with Fisher Score), 15NN	92.67	88.23
T2-30TS (30 best ratios with T-statistic), 3NN	89.25	90.44
T2-30TS (30 best ratios with T-statistic), 15NN	91.32	91.92
T3-59 Ratios (starting with all the 59 ratios), 3NN	94.62	82.35
T3-59 Ratios (starting with all the 59 ratios), 15NN	95.72	82.35
ANN, 3 selected ratios	90.35	92.65
ANN, 10 selected ratios	90.41	93.38

Selected inputs - GP as classifier	Accuracy	Sensitivity
Ratios from T1-30FS (30 best ratios with Fisher Score), 3NN	90.66	94.85
Ratios from T1-30FS (30 best ratios with Fisher Score), 15NN	92.61	92.65
Ratios from T2-30TS (30 best ratios with T-statistic), 3NN	91.48	92.65
Ratios from T2-30TS (30 best ratios with T-statistic), 15NN	92.34	92.65
Ratios from T3-59 Ratios, 3NN	93.04	92.65
Ratios from T3-59 Ratios, 15NN	92.30	93.38
Ratios used in ANN-3 Ratios	90.42	95.59
Ratios used in ANN-10 Ratios	91.87	92.65
Best ratios from GP as selector	93.15	94.85
59 ratios (without feature selection)	95.08	90.44

Classification measures using the complete test set, with inputs selected from the DE/KNN setup, and GP as a feature selection method

CONCLUSION

- The results show that the proposed selection method using GP stands out from the rest
- The use of GP as a classifier improves the results with respect to other classifier methods
- The study contributes to the field of business failure prediction by demonstrating the effectiveness of GP and evolutionary feature selection approaches.

THANK YOU