题目：Rough set and scatter search metaheuristic based feature selection for credit scoring

作者：Jue Wang，Abdel-Rahman Hedar，Shouyang Wang，Jian Ma

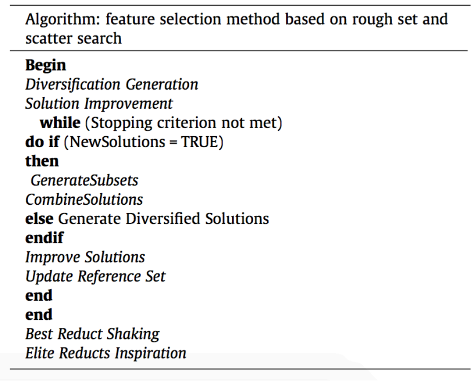
领域：信贷

核心创新点：RSFS(feature selection based on rough set and scatter search)特征工程

论文结构与实现方法：

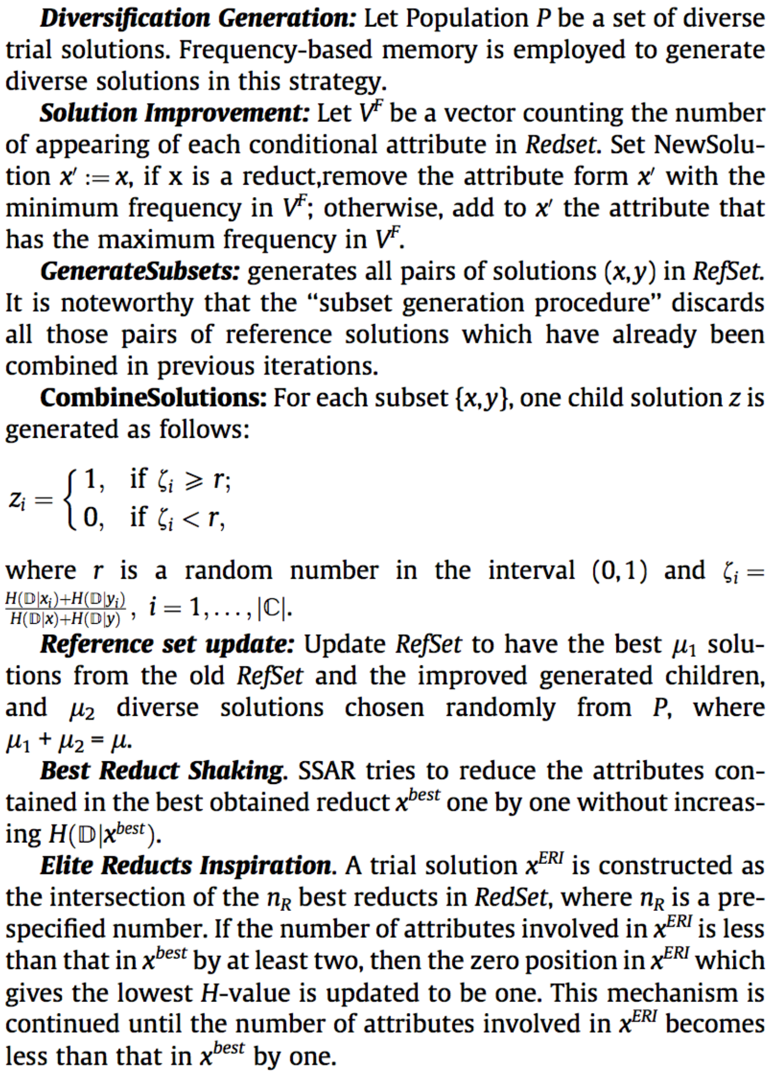
总述：为了处理大量信息冗余的信用数据集，提出了 RSFS 特征工程方法。在 RSFS 中，条件熵被认为是搜索最优解的启发式。选择 UCI 数据库中的两个信用数据集来描述包括神经网络模型，J48 决策树和 Logistic 回归在内的三个信用模型的 RSFS 的竞争性能。实验结果表明，与基本分类方法相比，RSFS 在节省计算成本和提高分类准确度方面具有优越的性能。

一、方法介绍(特征选择)——RSFS



Algorithm: feature selection method based on rough set and scatter search

```
Begin
Diversification Generation
Solution Improvement
   while (Stopping criterion not met)
do if (NewSolutions = TRUE)
then
 GenerateSubsets
CombineSolutions
else Generate Diversified Solutions
endif
Improve Solutions
Update Reference Set
end
end
Best Reduct Shaking
Elite Reducts Inspiration
```



Diversification Generation: Let Population *P* be a set of diverse trial solutions. Frequency-based memory is employed to generate diverse solutions in this strategy.

Solution Improvement: Let V^F be a vector counting the number of appearing of each conditional attribute in *Redset*. Set NewSolution $x' := x$, if x is a reduct,remove the attribute form x' with the minimum frequency in V^F; otherwise, add to x' the attribute that has the maximum frequency in V^F.

GenerateSubsets: generates all pairs of solutions (x, y) in *RefSet*. It is noteworthy that the “subset generation procedure” discards all those pairs of reference solutions which have already been combined in previous iterations.

CombineSolutions: For each subset $\{x, y\}$, one child solution *z* is generated as follows:

$$z_i = \begin{cases} 1, & \text{if } \zeta_i \geqslant r; \\ 0, & \text{if } \zeta_i < r, \end{cases}$$

where r is a random number in the interval $(0,1)$ and $\zeta_i = \frac{H(\mathbb{D}|x_i)+H(\mathbb{D}|y_i)}{H(\mathbb{D}|x)+H(\mathbb{D}|y)}$, $i = 1, \ldots, |\mathbb{C}|$.

Reference set update: Update *RefSet* to have the best μ_1 solutions from the old *RefSet* and the improved generated children, and μ_2 diverse solutions chosen randomly from *P*, where $\mu_1 + \mu_2 = \mu$.

Best Reduct Shaking. SSAR tries to reduce the attributes contained in the best obtained reduct x^{best} one by one without increasing $H(\mathbb{D}|x^{best})$.

Elite Reducts Inspiration. A trial solution x^{ERI} is constructed as the intersection of the n_R best reducts in *RedSet*, where n_R is a prespecified number. If the number of attributes involved in x^{ERI} is less than that in x^{best} by at least two, then the zero position in x^{ERI} which gives the lowest *H*-value is updated to be one. This mechanism is continued until the number of attributes involved in x^{ERI} becomes less than that in x^{best} by one.

1. 多样化生成

   产生用于生成不同 0/1 向量的 Glover 系统(SS，Scatter Search，散射搜索，某种人口进化算法)

2. 解决方法改进
3. 初始 RefSet
4. 找到最优特征子集
5. 强化程序改进最佳方案

二、实验验证

通过三种模型径向基函数（RBF），逻辑回归模型和 J48 决策树来比较应用 RSFS 和未应用该特征选择算法的差异。

三、数据集

机器学习数据库 UCI 存储库：

澳大利亚信贷数据库，案例 690，属性 14（6 个连续属性，8 个分类属性）

日本消费者信用卡申请批准，案例 664，属性 15

编辑日期：2017 年 10 月 7 日