Preprint of the paper:

Multi-Objective Particle Swarm Optimization for Feature Selection in Credit Scoring

by

Nikita Kozodoi, Stefan Lessmann

The final, revised paper is available from the publisher's website:

`https://link.springer.com/chapter/10.1007/978-3-030-66981-2_6`

Abstract

Credit scoring refers to the use of statistical models to support loan approval decisions. An ever-increasing availability of data on potential borrowers emphasizes the importance of feature selection for scoring models. Traditionally, feature selection has been viewed as a single-objective task. Recent research demonstrates the effectiveness of multi-objective approaches. We propose a novel multi-objective feature selection framework for credit scoring that extends previous work by taking into account data acquisition costs and employing a state-of-the-art particle swarm optimization algorithm for feature search. Our framework optimizes three fitness functions: the number of features, data acquisition costs and the AUC. Experiments on nine real-world credit scoring data sets demonstrate a highly competitive performance of the proposed framework.

Keywords: credit scoring, feature selection, multi-objective optimization, particle swarm optimization

MULTI-OBJECTIVE PARTICLE SWARM OPTIMIZATION FOR FEATURE SELECTION IN CREDIT SCORING

PREPRINT

Nikita Kozodoi*
Humboldt University of Berlin | Monedo
Berlin, Germany

Stefan Lessmann
Humboldt University of Berlin
Berlin, Germany

ABSTRACT

Credit scoring refers to the use of statistical models to support loan approval decisions. An ever-increasing availability of data on potential borrowers emphasizes the importance of feature selection for scoring models. Traditionally, feature selection has been viewed as a single-objective task. Recent research demonstrates the effectiveness of multi-objective approaches. We propose a novel multi-objective feature selection framework for credit scoring that extends previous work by taking into account data acquisition costs and employing a state-of-the-art particle swarm optimization algorithm for feature search. Our framework optimizes three fitness functions: the number of features, data acquisition costs and the AUC. Experiments on nine real-world credit scoring data sets demonstrate a highly competitive performance of the proposed framework.

***K*eywords** Credit scoring · Feature selection · Multi-objective optimization · Particle swarm optimization

1 Introduction

Financial institutions use credit scoring models to support loan approval decisions [10]. Due to the unprecedented availability of data on potential credit applicants and growing access of financial institutions to new data sources, the data used to train scoring models tend to be high-dimensional [6].

Feature selection aims at removing irrelevant features to improve the model performance, which is traditionally considered as a single-objective task [7]. In credit scoring, feature selection can be treated as a multi-objective problem with multiple goals. In addition to optimizing model performance, companies strive to reduce the number of features as public discourse and regulatory requirements are calling for comprehensible credit scoring models [9]. Furthermore, financial institutions often purchase data from external providers such as credit bureaus and banks in groups of features. This creates a need for a separate account for the data acquisition costs [13]. The conflicting nature of these objectives motivates us to consider feature selection as a multi-objective optimization problem.

Recent research has demonstrated the effectiveness of multi-objective feature selection in credit scoring using genetic algorithms (GA) [9]. However, previous studies have not considered a simultaneous optimization of the number and the cost of features. Moreover, techniques based on particle swarm optimization (PSO) have been recently shown to outperform GAs in other domains [15].

This paper proposes a multi-objective feature selection framework for credit scoring and makes two contributions. First, our framework addresses three distinct objectives: (i) the number of features, (ii) the cost of features and (iii) the model performance. Optimizing both the number and the cost of features is crucial as

*E-mail: *nikita.kozodoi@hu-berlin.de*

purchasing data from multiple sources introduces grouping that decorrelates the two objectives. Increasing the number of data providers incurs higher costs, whereas adding individual features does not affect costs if the new features are purchased in a group with the already included features. Second, we adjust a state-of-the-art PSO algorithm denoted as AgMOPSO to the feature selection context to improve the performance of the search algorithm.

2 Related Work

Standard techniques consider feature selection as a single-objective task [7]. Recent studies have shown the importance of accounting for multiple objectives such as model performance and cardinality of the feature set [3, 9]. A simple approach to account for multiple objectives is to aggregate them into one objective or introduce optimization constraints to a single-objective task. This requires additional information such as weights of objectives or budget conditions. In contrast, a truly multi-objective approach produces a set of non-dominated solutions – a Pareto frontier – where improving one objective is impossible without worsening at least one of the others. Provided with such a frontier, a decision-maker can examine the trade-off between the objectives depending on the context.

Previous research on multi-objective feature selection has employed evolutionary algorithms such as GA and PSO. Most studies use a non-dominated sorting genetic algorithm NSGA-II, which is a well-known algorithm for multi-objective optimization [8]. NSGA-III was proposed as a successor of NSGA-II to handle challenges of many-objective optimization [3]. The usage of GAs has also been recently challenged by the proposal of PSO techniques that demonstrate a superior performance [15, 16]. Research outside of the feature selection domain has also suggested other optimization techniques such as Gaussian processes [4].

Credit scoring is characterized by multiple conflicting objectives, such as model performance and comprehensibility [9, 13]. Nevertheless, research on multi-objective feature selection in credit scoring has been scarce. Maldonado et al. use support vector machines (SVM) to optimize performance while minimizing feature costs as a regularization penalty [12, 13]. Both methods are embedded in SVMs and output a single solution instead of a Pareto frontier. Kozodoi et al. argue for a model-agnostic multi-objective approach [9]. Their framework uses NSGA-II and is limited to two objectives, assuming the number of features to be indicative of both model comprehensibility and data acquisition costs.

We extend the previous work on feature selection in credit scoring by adapting a state-of-the-art PSO algorithm to perform the feature search and considering feature costs as a distinct objective. A common practice of purchasing data in groups of features reduces the correlation between the objectives and provides an opportunity for multi-criteria optimization. The number of features serves as a proxy for model comprehensibility and interpretability, whereas feature costs indicate the data acquisition costs faced by a financial institution.

3 Proposed Framework

We propose a multi-objective feature selection framework based on the external archive-guided MOPSO algorithm (AgMOPSO), which demonstrates superior performance compared to GAs in various optimization tasks [17]. PSO is a meta-heuristic method that solves an optimization problem by evolving a population of candidate solutions (i.e., particles) that iteratively navigate through the search space. AgMOPSO encompasses two stages: initialization and feature search.

The algorithm initializes with a random swarm of n particles. Each particle is represented by a real-valued vector of length k, where k is the number of features. The particle values are restricted to $[0, 1]$ and indicate the probability of a feature being selected. The initialization is followed by an iterative process of guiding the swarm towards new solutions using the immune-based and PSO-based search and evolving an archive that stores the non-dominated solutions.

The immune-based search produces new particles by applying genetic operators such as cloning, crossover and mutation to the existing particles [17]. The PSO-based search creates new particles by updating the values of each particle using a decomposition-based approach. The particle position in the k-dimensional feature space is adjusted by moving them towards the swarm leaders, i.e., particles that perform best in each of the three objectives [1].

Table 1: Credit Scoring Data Sets

Data set	Sample size	No. features	Default rate
australian	690	42	.44
german	1,000	61	.30
thomas	1,125	28	.26
hmeq	5,960	20	.20
cashbus	15,000	1,308	.10
lendingclub	43,344	206	.07
pakdd2010	50,000	373	.26
paipaidai	60,000	1,934	.07
gmsc	150,000	68	.07

After each search round, we evaluate solutions. We train a model that includes features corresponding to the rounded particle values and evaluate three fitness functions: (i) the number of features, (ii) feature acquisition costs and (iii) the model performance in terms of the area under the ROC curve (AUC). Based on the evaluated fitness, we store non-dominated solutions representing different feature subsets in the archive. Until the maximum size of the archive is reached, all non-dominated solutions are added to the archive. Once the archive is full, we calculate the crowding degree of new particles that indicates the density of surrounding solutions. If the new solution displays a better crowding degree than at least one archive solution, it replaces the most crowded solution in the archive.

4 Experimental Setup

4.1 Data

Table 1 displays the data sets used in the experiments. All data sets come from a retail credit scoring context. Data sets *australian* and *german* are part of the UCI Repository. Data sets *thomas* and *hmeq* are provided by [14] and [2]; *paipaidai* is collected from [11]. Data sets *pakdd*, *lendingclub* and *gmsc* are provided for data mining competitions on PAKDD and Kaggle platforms.

Each data set contains a binary target variable indicating whether a customer has repaid a loan and a set of features describing characteristics of the applicant, the loan and, in some cases, the applicant's previous loans. As illustrated in Table 1, the sample size and the number of features vary across the data sets, which allows us to test our feature selection framework in different conditions.

4.2 Experimental Setup

We consider a multi-criteria feature selection problem with three objectives: (i) the number of selected features, (ii) feature acquisition costs, (iii) the AUC. Each of the nine data sets is randomly partitioned into training (70%) and holdout sets (30%). We perform feature selection with AgMOPSO within four-fold cross-validation on the training set. The performance of the selected feature subsets is evaluated on the holdout set. To ensure robustness, the performance is aggregated over 20 modeling trials with different random data partitioning.

Since data on feature acquisition costs are not available in all data sets, we simulate costs similar to [15]. The cost of each feature is drawn from a Uniform distribution in the interval $[0, 1]$. To simulate feature groups, we introduce a cost-based grouping for categorical features. Each categorical feature is transformed with dummy encoding. Next, we assign acquisition costs to dummy features: if one dummy stemming from a specific categorical feature is selected, other dummies from the same feature can be included at no additional cost.

NSGA-II and NSGA-III with the same three objectives as AgMOPSO serve as benchmarks. The meta-parameters of the algorithms are tuned using grid search on a subset of training data. To ensure a fair comparison, the population size and the number of generations for NSGA-II and NSGA-III are set to the same values as for the AgMOPSO. We also use a full model with all features as a benchmark. L2-regularized logistic regression serves as a base classifier.

We use five evaluation metrics common in multi-objective optimization to reflect different characteristics of the evolved solution frontiers: (i) hypervolume (HV) is an overall performance metric that indicates the objective space covered by the solutions; (ii) overall non-dominated vector generation (ONVG) is the number of distinct non-dominated solutions; (iii) two-set coverage (TSC) reflects the convergence of the frontier; (iv) spacing (SPC) considers how evenly the solutions are distributed; (v) maximum spread (SPR) accounts for the solution spread. The calculation of the SPC and the SPR requires knowledge of the true Pareto frontier, which is difficult to estimate due to the high dimensionality of the feature space. We combine the non-dominated solutions across the three algorithms and 20 trials to form an adequate approximation of the true frontier.

We also compare the full model with the single best-performing solutions of the evolutionary algorithms that achieve the highest AUC compared to the other solutions on the evolved Pareto frontiers. The single solutions are compared in terms of the three considered objectives.

5 Results

Table 2 provides the experimental results. For each data set, we rank algorithms in the five multi-objective optimization metrics and report the mean ranks across the 20 trials. We also report the mean AUC, data acquisition costs and the number of features of the single solutions with the highest AUC.

Overall, AgMOPSO outperforms the GA-based benchmarks in three performance metrics, achieving the lowest average rank in the ONVG, the TSC and the HV. According to the Nemenyi test [5], differences in algorithm ranks are significant at a 5% level. The superior performance of AgMOPSO is mainly attributed to a higher cardinality and a better convergence of the evolved frontier compared to NSGA-II and NSGA-III. This is indicated by the best performance of AgMOPSO in the ONVG and the TSC on seven out of nine data sets.

In terms of the diversity of the evolved frontier, AgMOPSO does not outperform the benchmarks. In the feature selection application, diversity of solutions is of secondary importance as we are mainly interested in models from a subspace that covers the best-performing models. A more relevant indicator and an apparent strength of AgMOPSO compared to the competing algorithms is its better convergence to the true frontier, which indicates that it misses a smaller number of non-dominated solutions in the relevant search space.

The mean correlation between the number of features and data acquisition costs across the non-dominated solutions is $.7455$. This emphasizes the importance of a separate optimization of the two feature-based objectives.

Due to a large number of noisy features, the performance of the full model is suboptimal. Comparing the full model to the solutions of multi-objective algorithms with the highest AUC, we see that all three evolutionary algorithms identify a feature subset that achieves a higher AUC, lower costs and a smaller number of features on eight data sets. On average, AgMOPSO reduces data acquisition costs and the number of selected features by 58.77% and 76.20%, respectively. AgMOPSO also outperforms GA-based benchmarks in terms of the AUC on five data sets. At the same time, solutions of NSGA-III entail lower costs and a smaller number of features compared to AgMOPSO, which comes at the cost of a lower average AUC. These results indicate that AgMOPSO more effectively explores regions of the search space associated with a higher AUC.

Figure 1 illustrates the Pareto frontiers outputted by the three feature selection algorithms on one of the trials on *gmsc*. The figure demonstrates the ability of AgMOPSO to identify a larger number of non-dominated solutions in the search subspace with a high AUC compared to the GA-based benchmarks.

6 Discussion and Future Work

This paper proposes a multi-objective framework for feature selection in credit scoring using the AgMOPSO algorithm. We perform feature selection using three fitness functions reflecting relevant credit scoring objectives: the number of features, data acquisition costs, and model performance. The performance of our framework is assessed on nine real-world credit scoring data sets.

The results suggest that AgMOPSO is a highly competitive multi-objective feature selection framework, as indicated by standard quality criteria for multi-objective optimization. Compared to other evolutionary algorithms, AgMOPSO more effectively explores regions of the search space associated with a high model

Table 2: Comparing Performance of Feature Selection Methods

Data set	Algorithm	ONVG	TSC	SPC	SPR	HV	AUC	FC	NF
australian	AgMOPSO	**1.70**	**1.80**	2.25	**1.45**	**1.80**	**.9184**	1.96	9.35
	NSGA-II	2.05	1.93	2.00	2.10	2.15	.9170	1.95	8.15
	NSGA-III	2.10	2.05	**1.75**	2.15	2.05	.9162	**1.49**	**6.90**
	Full model			–			.6751	6.81	42
german	AgMOPSO	**1.20**	1.87	2.55	**1.35**	1.95	.7823	3.78	15.70
	NSGA-II	1.75	**1.82**	2.15	1.50	**1.40**	**.7824**	3.24	13.65
	NSGA-III	2.90	2.07	**1.30**	2.70	2.65	.7723	**1.95**	**9.05**
	Full model			–			.5806	9.97	61
thomas	AgMOPSO	1.65	**1.75**	2.05	**1.75**	1.95	.6368	1.47	4.50
	NSGA-II	1.70	**1.75**	2.05	1.80	**1.80**	.6363	1.36	3.70
	NSGA-III	**1.55**	**1.75**	**1.50**	1.95	1.85	.6364	**1.00**	**3.30**
	Full model			–			**.6375**	6.81	28
hmeq	AgMOPSO	**1.20**	**1.77**	2.45	**1.45**	**1.45**	**.7660**	3.23	9.45
	NSGA-II	1.95	1.88	2.15	1.75	2.25	.7651	3.14	9.10
	NSGA-III	2.60	2.02	**1.40**	2.60	2.30	.7604	**2.32**	**5.85**
	Full model			–			.5730	5.76	20
cashbus	AgMOPSO	**1.85**	**1.72**	1.70	2.25	2.30	.6450	**3.52**	**10.55**
	NSGA-II	1.95	2.00	2.90	**1.05**	2.30	.6426	14.64	44.95
	NSGA-III	2.15	2.08	**1.40**	2.70	**1.40**	**.6606**	8.22	27.50
	Full model			–			.5233	321.10	1308
lendingclub	AgMOPSO	**1.60**	1.95	1.90	1.70	2.05	.6169	1.96	20.25
	NSGA-II	1.80	2.12	2.30	**1.50**	2.60	.6149	2.08	18.60
	NSGA-III	2.40	**1.75**	**1.80**	2.25	**1.35**	**.6176**	**1.81**	**16.75**
	Full model			–			.5725	6.24	206
pakdd2010	AgMOPSO	**1.45**	**1.75**	2.15	2.05	**1.00**	**.6254**	8.10	115.55
	NSGA-II	1.95	2.28	2.75	**1.00**	2.00	.6252	9.47	252.05
	NSGA-III	2.55	1.95	**1.10**	2.95	3.00	.6220	**7.65**	**69.65**
	Full model			–			.5748	14.58	373
paipaidai	AgMOPSO	2.85	**1.80**	1.70	2.50	1.85	.6639	5.84	**36.15**
	NSGA-II	1.60	2.03	2.85	**1.00**	**1.40**	.6727	8.88	245.40
	NSGA-III	**1.50**	1.92	**1.45**	2.40	2.75	**.6802**	**5.82**	76.35
	Full model			–			.4956	57.89	1934
gmsc	AgMOPSO	**1.00**	**1.72**	2.45	**1.50**	**1.35**	**.8603**	4.72	25.50
	NSGA-II	1.95	1.90	2.20	1.70	1.90	.8602	4.71	23.15
	NSGA-III	3.00	2.30	**1.35**	2.65	2.75	.8556	**3.52**	**15.75**
	Full model			–			.6437	4.85	68

Note: ONVG = overall non-dominated vector generation, TSC = two-set coverage, SPC = spacing, SPR = maximum spread, HV = hypervolume, AUC = area under the ROC curve, FC = feature acquisition costs, NF = number of selected features.

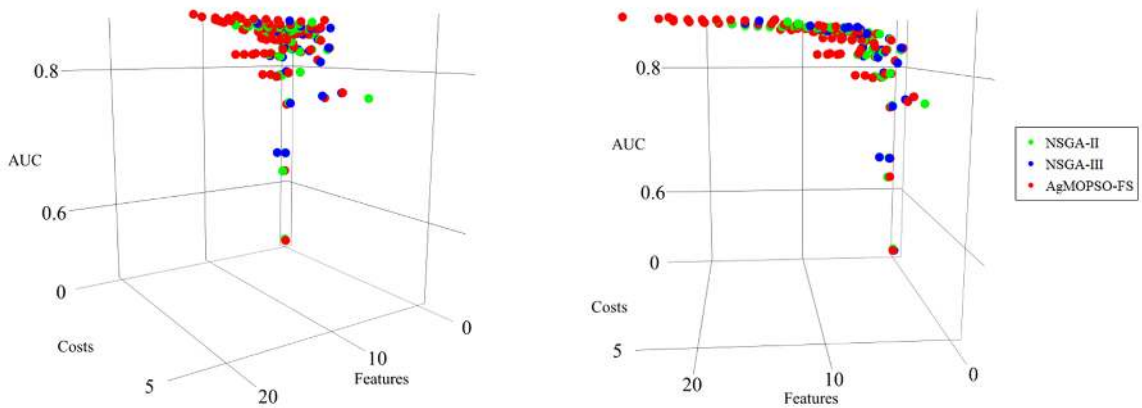

Figure 1: Pareto frontiers for *gmsc* depicted from two angles.

performance, while also substantially reducing the number of features and the data acquisition costs compared to a model using all features.

In future studies, we plan to conduct a more in-depth analysis of AgMOPOSO. It would be interesting to compare results with the solutions evolved by two-objective feature selection algorithms that ignore data acquisition costs. Analysis of the impact of correlation between the objectives on the algorithm performance could also shed more light on conditions in which the number and the cost of features should be considered as separate objectives. In addition, computing the running times and the number of generations before convergence would contribute a new angle to compare feature selection algorithms.

AgMOPOSO has a wide set of meta-parameters, which poses an opportunity for a systematic sensitivity analysis that could provide deeper insights into the appropriate values. For instance, the diversity of the evolved solutions could be improved by adjusting the crossover and mutation operations within the search. Using different base learners could help evaluate gains given a model with a built-in feature selection mechanism (e.g., L1-regularized logistic regression).

Our multi-objective feature selection framework could be extended to other application areas, such as fraud detection or churn prediction. For both of these applications, customer data are typically gathered from different sources and therefore provides opportunities for group-based cost optimization.

References

[1] Al Moubayed, N., Petrovski, A., McCall, J.: D2MOPSO: MOPSO based on decomposition and dominance with archiving using crowding distance in objective and solution spaces. Evolutionary Computation **22**(1), 47–77 (2014)

[2] Baesens, B., Roesch, D., Scheule, H.: Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley & Sons (2016)

[3] Bidgoli, A.A., Ebrahimpour-Komleh, H., Rahnamayan, S.: A many-objective feature selection algorithm for multi-label classification based on computational complexity of features. In: 2019 14th International Conference on Computer Science & Education (ICCSE). pp. 85–91. IEEE (2019)

[4] Bradford, E., Schweidtmann, A.M., Lapkin, A.: Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm. Journal of Global Optimization **71**(2), 407–438 (2018)

[5] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)

[6] Gambacorta, L., Huang, Y., Qiu, H., Wang, J.: How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm (2019), Working paper, Bank for International Settlements

[7] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature extraction: Foundations and applications, vol. 207. Springer (2008)

[8] Hamdani, T.M., Won, J.M., Alimi, A.M., Karray, F.: Multi-objective feature selection with NSGA II. In: International Conference on Adaptive and Natural Computing Algorithms. pp. 240–247. Springer (2007)

[9] Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., Baesens, B.: A multi-objective approach for profit-driven feature selection in credit scoring. Decision Support Systems **120**, 106–117 (2019)

[10] Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research **247**(1), 124–136 (2015)

[11] Li, H., Zhang, Y., Zhang, N.: Evaluating the well-qualified borrowers from PaiPaiDai. Procedia Computer Science **122**, 775–779 (2017)

[12] Maldonado, S., Flores, Á., Verbraken, T., Baesens, B., Weber, R.: Profit-based feature selection using support vector machines – general framework and an application for customer retention. Applied Soft Computing **35**, 740–748 (2015)

[13] Maldonado, S., Pérez, J., Bravo, C.: Cost-based feature selection for support vector machines: An application in credit scoring. European Journal of Operational Research **261**(2), 656–665 (2017)

[14] Thomas, L., Crook, J., Edelman, D.: Credit scoring and its applications. SIAM (2017)

[15] Zhang, Y., Gong, D.w., Cheng, J.: Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics **14**(1), 64–75 (2015)

[16] Zhang, Y., Gong, D.w., Sun, X.y., Guo, Y.n.: A PSO-based multi-objective multi-label feature selection method in classification. Scientific Reports **7**(1), 1–12 (2017)

[17] Zhu, Q., Lin, Q., Chen, W., Wong, K.C., Coello, C.A.C., Li, J., Chen, J., Zhang, J.: An external archive-guided multiobjective particle swarm optimization algorithm. IEEE Transactions on Cybernetics **47**(9), 2794–2808 (2017)