Multidimensional Time Series Feature Engineering by Hybrid Evolutionary Approach

Piotr Lipinski
Computational Intelligence Research Group, Institute of Computer Science, University of Wroclaw
Wroclaw, Poland
piotr.lipinski@cs.uni.wroc.pl

Krzysztof Michalak
Department of Information Technologies, Institute of Business Informatics, Wroclaw University of Economics
Wroclaw, Poland
krzysztof.michalak@ue.wroc.pl

ABSTRACT

This paper proposes a hybrid evolutionary approach to feature engineering for mining frequent patterns from multidimensional financial ultra-high frequency time series. Experiments performed on real-world data from the London Stock Exchange Rebuilt Order Book database confirms that the evolutionary algorithm is capable of improving significantly the results of classification.

ACM Reference Format:
Piotr Lipinski and Krzysztof Michalak. 2019. Multidimensional Time Series Feature Engineering by Hybrid Evolutionary Approach. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3326795

1 INTRODUCTION

This paper was inspired by the studies on computational intelligence algorithms for non-linear dimensionality reduction, especially manifold learning, which gain more and more popularity in recent years, among the others, the studies of Saul and Roweis on LLE, Scholdkopf, Smol and Muller on KPCA, Hinton and Roweis on SNE, van der Maaten and Hinton on t-SNE and their extensions. Although the purpose of these studies usually was mining or visualization of high-dimensional data, it is interesting to use these techniques to discover dependencies between the coordinates of the solution vector and to reduce the search space in evolutionary algorithms.

The proposed approaches strive to determine a certain subspace of the search space that probably contains the optimal solution and to search through it, instead of the entire search space. In order to determine the subspace, the current population is studied, considered as a data sample of the search space from the neighborhood of the optimal solution, and a new reduced search space (isomorphic to a certain subspace of the original search space) is determined using a dimensionality reduction technique applied to the data sample.

The hybrid evolutionary approach was applied to feature engineering for mining frequent patterns from multidimensional financial ultra-high frequency time series, based on real-world data from the London Stock Exchange Rebuilt Order Book database extending the approach proposed in [2].

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07.
https://doi.org/10.1145/3319619.3326795

2 THE DEELLE ALGORITHM

The Differential Evolution Enhanced by Locally Linear Embeddings (DEELLE) algorithm [3], is an extension of Differential Evolution (DE) [1] with the dimensionality reduction based on Locally Linear Embeddings (LLE) [5], which focuses on determining a non-linear transformation of the original search space into a reduced search space of lower dimensionality with preserving some local linearities.

3 FEATURE ENGINEERING WITH DEELLE

This paper proposes an evolutionary approach to parameterization of frequent patterns in financial ultra-high frequency time series, where input data describe Limit Order Books (LOBs) and target data indicate significant changes in the price of the financial asset under study. First, LOBs are encoded in a feature-based data representation. Second, a binary SVM classifier is trained to predict whether a particular LOB leads to a significant change in the stock price in a few successive time instants, or not. Third, an EA is used to find the optimal parameterization of the feature-based data representation, so that the performance of the classifier was better. The details of the problem were described in [4].

LOBs are encoded in the feature-based data representation using the Gaussian Density Filters (GDF) representation, introduced in [4]. It uses a predefined filter bank containing a number K of filters $f_1, f_2, \ldots, f_K$, where each filter f_k is a function $f_k : \mathbb{R} \to \mathbb{R}$ defined by a Gaussian curve with the parameters μ_k and σ_k, for $k = 1, 2, \ldots, K$. In the GDF representation, each order queue is represented by a feature vector $\mathbf{q} \in \mathbb{R}^K$, where each feature q_k corresponds to the similarity of the order queue to the k-th Gaussian Density Filter f_k measured in the following way:

$$\begin{aligned} q_k &= \sum_{i=1}^{N} \frac{v_i}{v_{total}} - \int_{p_{i-1}}^{p_i} f_k(x)dx \qquad (1) \\ &= \sum_{i=1}^{N} \frac{v_i}{v_{total}} - (\mathrm{CDF}(p_i; \mu_k, \sigma_k) - \mathrm{CDF}(p_i; \mu_k, \sigma_k)), \end{aligned}$$

for $k = 1, 2, \ldots, K$, where $\mathrm{CDF}(x; \mu_k, \sigma_k)$ denotes the value of the Gaussian cumulative distribution function with parameters μ_k and σ_k at point x, N denotes the length of the order queue, and (p_i, v_i) denote the pairs of the order queue.

The initial filter bank used in encoding the buy order queue contains $K = 50$ filters defined by $\mu_k = -k$ and $\sigma_k = 1$ and the

Algorithm 1 Differential Evolution Enhanced by Locally Linear Embeddings (DEELLE)

$\mathcal{P}_0$ = Random-Population(N)
Population-Evaluation($\mathcal{P}_0, F$)
$t = 0$
while not Termination-Condition($\mathcal{P}_t$) **do**
  for all $\mathbf{x} \in \mathcal{P}_t$ **do**
    pick randomly distinct $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ from $\mathcal{P}_t \setminus \{\mathbf{x}\}$
    $\mathbf{v} = \mathbf{x}_1 + \alpha \cdot (\mathbf{x}_2 - \mathbf{x}_3)$
    $\mathbf{u}$ = Binomial-Recombination($\mathbf{v}, \mathbf{x}$)
    if $F(\mathbf{x}) \leq F(\mathbf{u})$ **then**
      $\mathbf{u}$ replaces $\mathbf{x}$ in $\mathcal{P}_{t+1}$
    end if
  end for
  Population-Evaluation($\mathcal{P}_{t+1}, F$)
  $t = t + 1$
  if Subevolution-Starting-Condition() **then**
    Search-Space-Reduction()
    $\mathcal{R}_0$ = Population-Reduction($\mathcal{P}_t$)
    $s = 0$;
    while not Subevolution-Termination-Condition($\mathcal{R}_s$) **do**
      for all $\mathbf{x} \in \mathcal{R}_s$ **do**
        pick randomly distinct $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ from $\mathcal{R}_s \setminus \{\mathbf{x}\}$
        $\mathbf{v} = \mathbf{x}_1 + \alpha \cdot (\mathbf{x}_2 - \mathbf{x}_3)$
        $\mathbf{u}$ = Binomial-Recombination($\mathbf{v}, \mathbf{x}$)
        if $F(\mathbf{x}) \leq F(\mathbf{u})$ **then**
          $\mathbf{u}$ replaces $\mathbf{x}$ in $\mathcal{R}_{s+1}$
        end if
      end for
      Reduced-Population-Evaluation($\mathcal{R}_{s+1}, F$)
      $s = s + 1$
    end while
    Search-Space-Restoring()
    $\mathcal{P}_t$ = Population-Restoring($\mathcal{R}_{s-1}$)
  end if
end while

initial filter bank used in encoding the sell order queue contains $K = 50$ filters defined by $\mu_k = k$ and $\sigma_k = 1$ (as proposed in [4]). The parameters of the both initial filter banks are subject of evolutionary optimization aiming at improving the performance of the classifier trained with the feature-based data representation.

For a given number K of filters, let μ_k^B and σ_k^B (for $k = 1, 2, \ldots, K$) denote the parameters from the filter bank for encoding the buy order queue and let μ_k^S and σ_k^S denote the parameters from the filter bank for encoding the sell order queue. The objective function $\mathcal{F}(\mathbf{x})$ for a candidate solution is AUC of the binary SVM classifier constructed using the feature-based data representation with the filter banks containing filters with the parameters $\mu_k^B, \mu_k^S, \sigma_k^B, \sigma_k^S$ defined by the candidate solution $\mathbf{x}$.

Finding optimal parameters μ_k^B,σ_k^B, μ_k^S, $\sigma_k^S$of the filter bank constitutes a difficult optimization problem with a high dimensional search space $\Omega = \mathbb{R}^{2 \cdot 2 \cdot K}$. Although the search space is high dimensional, there may exist some dependencies between the parameters, which may reduce the dimensionality of the search space, at least in a certain neighborhood of the optimal solution.

Table 1: Summary of results

ISIN	Baseline AUC	Optimized AUC	Difference
GB0005405286	0,6973	0,7566	0,0593
GB00B16GWD56	0,7147	0,7833	0,0686
GB0007980591	0,7724	0,8368	0,0644
GB0009252882	0,7397	0,7915	0,0518
GB00B03MLX29	0,7810	0,8307	0,0497
GB0002875804	0,7073	0,7860	0,0787
GB00B03MM408	0,8260	0,8857	0,0597
GB0002374006	0,7115	0,7970	0,0855
GB0008762899	0,6396	0,6869	0,0473
GB0000566504	0,6521	0,7058	0,0537
GB0009895292	0,7315	0,7775	0,0460
GB0031348658	0,6715	0,7419	0,0704
GB0007188757	0,6712	0,7309	0,0597
GB0008706128	0,6340	0,6581	0,0241
GB00B10RZP78	0,7298	0,7629	0,0331
GB0008847096	0,7492	0,7868	0,0376
GB0004835483	0,7036	0,7563	0,0527
GB0004082847	0,6690	0,7473	0,0783
GB00B24CGK77	0,6844	0,7463	0,0619
GB0007099541	0,6420	0,6878	0,0458

The DEELLE algorithm was used to solve the optimization problem. Experiments concerned financial ultra-high frequency time series from the London Stock Exchange Rebuild Order Book (LSEROB) database for 20 financial assets that are the top 20 components of the FTSE100 index with the highest weights over the period between the September 1, 2013 and September 15, 2013 (10 trading days). Using DEELLE led to improving the AUC of the binary SVM classifier, as presented in Table 1.

4 CONCLUSIONS AND PERSPECTIVES

This paper proposed a hybrid approach to parameterization of patterns in financial time series based on the DEELLE algorithm that enhances differential evolution with nonlinear dimensionality reduction of the search space. Experiments confirmed that the proposed approach was capable of optimizing the overall performance of the approach to predict significant changes in the price on the basis of the LOB data.

ACKNOWLEDGMENT

Calculations have been carried out using resources provided by Wroclaw Centre for Networking and Supercomputing (http://wcss.pl), grant no. 405.

REFERENCES

[1] S. Das and P. Suganthan. [n. d.]. Differential Evolution: A Survey of the State-of-the-Art. 15, 1 ([n. d.]), 4–31.
[2] Martin D. Gould, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. [n. d.]. Limit order books. *Quantitative Finance* 13, 11 ([n. d.]), 1709–1742. https://doi.org/10.1080/14697688.2013.803148
[3] Piotr Lipinski. 2014. Training Complex Decision Support Systems with Differential Evolution Enhanced by Locally Linear Embedding. In *Applications of Evolutionary Computation, EvoWorkshops 2014*. Lecture Notes in Computer Science, Vol. 8602. Springer, 125–137.
[4] Piotr Lipinski. 2017. Optimization of Representation for Extracting Knowledge from Ultra-High Frequency Time Series. In *Proceedings of the IEEE International Conference on Evolutionary Computing (CEC)*. IEEE.
[5] S. Roweis and L. Saul. [n. d.]. Nonlinear Dimensionality Reduction by Locally Linear Embedding. 290, 5500 ([n. d.]), 2323–2326.