Inspecting the Latent Space of Stock Market Data with Genetic Programming

Sungjoo Ha
School of Computer Science & Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu,
Seoul, 151-744 Korea
shurain@soar.snu.ac.kr

Sangyeop Lee
School of Computer Science & Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu,
Seoul, 151-744 Korea
leesy714@soar.snu.ac.kr

Byung-Ro Moon
School of Computer Science & Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu,
Seoul, 151-744 Korea
moon@snu.ac.kr

ABSTRACT

We suggest a method of inspecting the latent space of stock market data using genetic programming. Given black box patterns and ⟨stock, day⟩ tuples a relation matrix is constructed. Applying a low-rank matrix factorization technique to the relation matrix induces a latent vector space. By manipulating the latent vector representations of black box patterns, the geometry of the latent space can be examined. Genetic programming constructs a tree representation corresponding to an arbitrary latent vector representation, allowing us to interpret the result of the inspection.

CCS Concepts

•**Computing methodologies** → **Non-negative matrix factorization; Genetic programming;**

Keywords

matrix factorization; latent space models; technical patterns; genetic programming;

1. MOVITATION

Suppose we have human expert traders trading in a stock market. It is desirable to specify how they make decision into set of rules or classifiers to build an automatic trading system. Furthermore, systematic comparison of multiple experts can yield new insights. We can perform such comparison by inspecting the latent space induced by low-rank matrix factorization.

2. PROBLEM STATEMENT

A pattern is defined to be a classifier that yields a true/false result for a given ⟨stock, date⟩ tuple. A black box pattern is a pattern that we do not necessarily know the decision rules it applies. Given multiple black box patterns and their results, we can create a relation matrix whose rows and columns correspond to patterns and ⟨stock, date⟩ tuples.

We apply a non-negative matrix factorization (NMF) [2] to the given relation matrix. This creates two rank-k matrices each corresponding to patterns and ⟨stock, date⟩ tuples. A row corresponding to a pattern is understood as a new vector representation of the pattern. Similarly, a column corresponding to a ⟨stock, date⟩ tuple can be seen as a new vector representation of the tuple. Since these new representations are embedded in a k-dimensional vector space, we can perform usual vector operations on them. This allows us to systematically compare different patterns. For example, we can easily perform clustering on these new representations and find a new pattern that corresponds to the center of a cluster.

While low-rank matrix factorization techniques allow us to

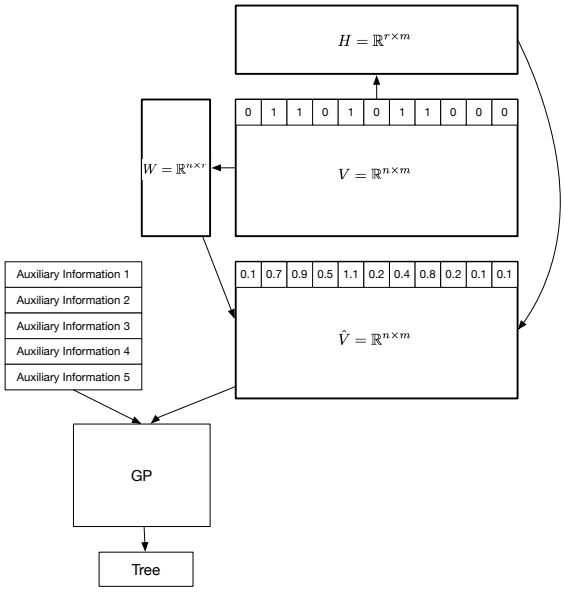

Figure 1: NMF is performed on a relation matrix between patterns and ⟨stock, date⟩ tuples to induce a latent space. An arbitrary point in the latent pattern space can form a new row of $\hat{V}$ which corresponds to the behavior of the hypothetical pattern. We can recover a tree representation of the pattern using auxiliary information and GP.

GECCO '16 July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.

DOI: http://dx.doi.org/10.1145/2908961.2909004

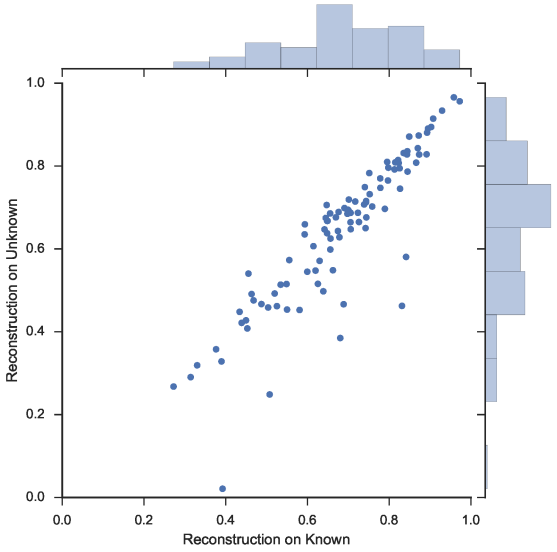

Figure 2: The joint plot of reconstruction accuracy on known and unknown data. High correlation between two results indicates that a tree that mimics the original behavior tends to generalize to unseen data as well.

create new representations, an arbitrary latent vector representation does not have a correponding decision rule. Therefore we cannot use such representations directly to predict the future behavior of said hypothetical patterns. But we can use auxiliary information to build trees corresponding to these patterns using genetic programming (GP). An arbitrary latent vector is multiplied to ⟨stock, date⟩ tuple matrix to create a behavior vector. GP finds a tree that mimics the behavior of this particular vector.

3. EXPERIMENTAL RESULTS

To test our approach, we created pattern trees using GP whose objective was to find attractive patterns [1]. Having such white box patterns allows us the make quantitative comparison of the results. We used Korean stock market data from 2013 to 2014 for the experiments.

3.1 Recovering Original Patterns

First, we show that it is possible to create patterns that mimic the behavior of original patterns and that they generalize to unseen data. Figure 2 depicts the relationship between reproducibility and generalizability. Trees that were able to accurately reproduce the behavior of original trees were also able to generalize to unseen data. This establishes that trees recovered by GP models the behavior of black box patterns. It is noteworthy to point out that newly created trees were different from the original trees in terms of their constituent nodes.

3.2 Cluster Centers

To showcase the ability to inspect the latent space, we study the clustering problem. We performed k-means clustering algorithm to the latent vector representation of patterns. After identifying clusters, each cluster center was computed by taking the average of patterns in the same cluster. By applying our approach, tree representations corresponding to cluster centers were created.

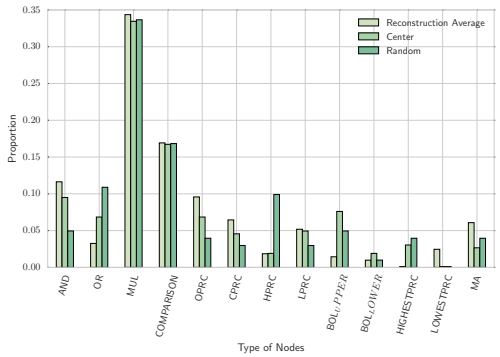

Figure 3: Comparison of the composition of trees. A tree consists of different type of nodes and their proportion varies from tree to tree. Reconstructed average corresponds to the average proportion of nodes for reconstructed trees of the same cluster. Center corresponds to a hypothetical pattern corresponding to the cluster center. Notice that the center is more similar to the reconstructed average than to a random pattern tree.

We compared the proportion of different node types to confirm that cluster centers are indeed similar to other patterns in the same cluster. The average KL divergence between reconstruction trees and the cluster center is 0.1575 and the average KL divergence between reconstruction trees and a random tree is 0.2178. Figure 3 illustrates a typical case of such a comparison.

4. CONCLUSIONS

We established that it is possible to combine low-rank matrix factorization and genetic programming to inspect the latent space of stock market data. This approach does not necessarily have to be limited to stock markets. In principle, it can be applied to arbitrary relation matrix together with auxiliary data. We hope to apply this method in other domains in the future.

5. ACKNOWLEDGMENTS

This work is the result of a commercial project conducted at Optus Investment Inc. The work was also partly supported by the Engineering Research Center of Excellence Program of Korea Ministry of Science, ICT & Future Planning(MSIP) / National Research Foundation of Korea(NRF) (Grant NRF-2008-0062609). The ICT at Seoul National University provided some research facilities for this study.

References

[1] S. Ha and B. R. Moon. Fast Knowledge Discovery in Time Series with GPGPU on Genetic Programming. *Genetic and Evolutionary Computation Conference*, pages 1159–1166, 2015.

[2] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems 13*, pages 556–562. 2001.