Visualization of Financial Data Space

Abstract

Today, data bases become increasingly large as well as increasingly complex. Visualization of high dimensional data is an important problem in many different domains and especially in Finance. Visualization of financial data and a good representation of the data space is useful in many financial applications.

The main problem of visualization of high dimensional data concerns the data representation in 2D or 3D with minimal loss of information. Furthermore, data visualization techniques are using dimensionality reduction methods in order to get a 2D or 3D representation of the data. The dimensionality reduction aim is to preserve as much of the significant structure of the high-dimensional data as possible in the low-dimensional map.

Here we present the first implementation of t-Distributed Stochastic Neighbor Embedding (t-SNE) optimization method for the visualization and the representation of the financial data space. In order to get a good and relevant representation of the data space, we coupled the t-SNE algorithm with an optimization engine for feature selection. To test the algorithm, we used two bankruptcy data bases. The first data base contains 250 companies (143 healthy and 107 bankrupt), were each company is characterized by 7 features. The second data base contains 500 companies (250 healthy and 250 bankrupt), were each company is characterized by 40 features of financial ratios.

The newly t-SNE optimization algorithm produce a 2D representations of the data bases. The 2D representations were evaluated by standard parameters. The algorithm captured much of the local information of the high-dimensional data very well, while also revealing global information such as the financial data space, which clearly shows visual separation of the data to the correct clusters.

Introduction

Today, data bases become increasingly large as well as increasingly complex. Visualization of high dimensional data is an important problem in many different domains and especially in Finance. Visualization of financial data old some promise for representing data in an easily understandable format.

The main problem of visualization of high dimensional data concerns the data representation in 2D or 3D with minimal loss of information. Furthermore, data visualization techniques are using dimensionality reduction methods in order to get a 2D or 3D representation of the data. The dimensionality reduction aim is to preserve as much of the significant structure of the high-dimensional data as possible in the low-dimensional map.

A common technique for dimension reduction and visualization is Principal Component Analysis (PCA)(1). PCA reduces the dimensionality of a data set, while retaining as much as possible, its original variance by transforming the original features into a new set of orthogonal features. Another common method is the Self-Organized Map (SOM). SOM reduces the dimensions of data through the use of self-organizing neural networks, by clustering together similar objects.  The latest methods in this area is the t-Distributed Stochastic Neighbor Embedding (t-SNE) (2) which reduces the dimensions of data by using probability distribution to locate similar objects and dissimilar objects. t-SNE is capable of retaining the local structure of the data while also revealing some important global information, such as clusters.

In the area of finance and economics dimension reduction methods have previously been employed for visual and analysis financial data. SOM have been applied in illustrating the performance of European banks(3), bankruptcy visualization(4), and for decomposing and identifying temporal structural changes in macro-financial data before, during and after the global financial crisis of 2007-2009(5).

Here we present the first implementation of t-SNE optimization method for the visualization and the representation of the financial data space. In order to get a good and relevant representation of the space and to take the advantage that t-SNE can discover global information such as clusters, we coupled the t-SNE algorithm with a genetic algorithm as optimization engine for feature selection. By using an optimization engine for feature selection we can obtain not only 2D representation of the data set but also a representation that is separated to the different clusters. To test the algorithm, we use two bankruptcy data set that were retrieved from the web.

METHODS

Data sets

Genetic Algorithms (GA)

Genetic algorithms[REF] belongs to the family of Evolutionary Algorithms. The method iteratively produces a population of solutions so that each population contains solutions which are better than those found in the previous population. This is done by first generating a random population of solutions where each solution is mapped into a "chromosome", each corresponding to a unique solution to the optimization problem and evaluating each chromosome based on the performances on the fitness function. finally, by letting the chromosomes that correspond to the better solution transfer their "genetic information" (i.e., features) to subsequent generations using a set of genetic operators such as selection of the fittest, mutation and cross-over.

This process iterates multiple times (generations) until no improvement in the fitness function is observed or until the number of pre-defined generation has been exhausted.

**Statistical Parameters for Evaluation**

In all cases the t-SNE optimization algorithm was evaluated using two parameters

1. The 1-Nearest Neighbor (1NN) classifier accuracy on the low-dimensional representation. The percent of instance that the closest neighbor in the 2D representation is from the same class.
2. The TrustWorthiness (TW) of the low-dimensional embedding.

**Results and discussion**

**Conclusions**

**References**

1. Jolliffe IT. Principal Component Analysis and Factor Analysis. Principal Component Analysis. Springer Series in Statistics: Springer New York; 2002.

2. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579-605.

3. Sarlin P, Eklund T. Financial performance analysis of European banks using a fuzzified self-organizing map. International Journal of Knowledge-Based and Intelligent Engineering Systems. 2013;17(3):223-34.

4. López Iturriaga FJ, Sanz IP. Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. Expert Systems with Applications. 2015;42(6):2857-69.

5. Sarlin P. Decomposing the global financial crisis: A Self-Organizing Time Map. Pattern Recognition Letters. 2013;34(14):1701-9.