T-SNE Classifier

# Abstract

This paper describes our utilization of the T-SNE dimensionality reduction algorithm as a classifier.

# Introduction

Today, databases become increasingly large as well as increasingly complex. Visualization of high dimensional data is an important problem in many different domains and especially in drug design. Visualization of chemical data and a good representation of the chemical space is useful in many chemoinformatic and drug design applications including the selection of compounds to synthesis, the selection of compounds for biological evaluation, selection of subsets for the design of information-rich compound libraries, and even for the development of reliable QSAR models. 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.

An additional domain in which visualization of high dimensional data is the financial domain where publicly traded firms are required by law to provide a plethora of financial indicators and leave it to the investors to make sense of that data. Observing the data in a human understandable two dimensional (or even three dimensional) space allows human analysts to recognize structure and trends in the constitution and robustness of the publicly traded firms and avoid the firms on the path to bankruptcy.

Here we present the first implementation of t-Distributed Stochastic Neighbor Embedding (t-SNE) [1] method for the visualization and the representation of the chemical space. In order to get a good representation of the chemical space we coupled the t-SNE algorithm with an optimization engine for feature selection. To test the algorithm, we used two data bases (1) The Comprehensive Medicinal Chemistry (CMC) database. This database contains 4,855 pharmaceutical compounds classified into 105 different biological indications, were each compound is characterized by 39 calculated descriptors. (2) The Bitter database. This database contains 1527 non-bitter compounds and 547 bitter compounds, were each compound is characterized by 19 calculated descriptors.

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

Visualization is especially beneficial in the concept exploration phase of the systems engineering process. Even though the raw data is available the information required for exact calculations for each proposed concept has to be acquired in a costly and time consuming manner. Reducing the data to two dimensions and visualizing it can provide powerful intuition into the behavior of the data and hint which alternatives are more likely to provide successful solutions.

# Definitions

**Quality**: a 0 to 1 measure of the classification. 0 means total classification failure and 1 means perfect classification success. Quality measures the success of the classification process by relying on labels.

**Trustworthiness**: a measure of how close points remained after the dimensionality reduction as described in [2]. Trustworthiness measures the success of the clustering process and is agnostic to labels.

**Feature Analysis**: a measure of classification success for each label in the dataset. 0 means that no point of that label has a nearest neighbor of the same label and 1 means that all points of that label have a nearest neighbor of the same label. Feature analysis measures the relevance of the success of the classification process for the specific label.

# The Algorithm

1. Randomly determine which features will be included in the dataset:
   1. Generate a uniformly distributed random number between 1 and 2n-1 where n is the number of features in the database.
2. Using Genetic Algorithm mutate the current number:
   1. Either add up to X features (by turning their bits from 0 to 1)
   2. Or remove up to X features (by turning their bits from 1 to 0)
   3. Or remove and add.
3. Use the number to generate the reduced dataset:
4. Execute the T-SNE algorithm reducing the dataset to 2 dimensions
5. Calculate result quality:
   1. For each point in the 2 dimensional dataset find the closest point. If they have the same label increment the quality measure by
   2. Sum up the partial quality measures to a total quality measure.
6. Repeat quality calculation by using majority vote from 3 neighbors.
7. Repeat quality calculation by using majority vote from 6 neighbors.
8. Calculate trustworthiness.
9. Calculate feature analysis.
10. If the quality if higher than before repeat step 2 with the new number. Otherwise repeat step 2 with the previous number.

# Results

The algorithm was executed multiple times for several databases:

## CMC

A database describing SOMETHING

The result images are in the CMC folder and the summary is in the CMC.csv file.

The CMC database is the only database with more than 2 labels. The quality measure of the results is in the range of [0.113, 0.215] with the 3 nearest neighbors quality measure significantly lower and 6 nearest neighbors quality measure even lower.

In addition it seems that the trustworthiness of results drop as the quality measure rises. The following graph shows quality and trustworthiness measure of points sorted by quality measure.

## Bitter

The database describing SOMETHING ELSE

The samples are labeled with 0 and 1 to represent SOMETHING

In certain datasets the quality measure is more than 90% - meaning that most of the points had a closest point with the same label and therefore the classifier was successful.

The negative correlation between trustworthiness and quality persists in this database.

The following graph shows quality and trustworthiness measure of points sorted by quality measure.

## Financial Ratios

The financial ratios database represents some institutional financial indicators (such as debt to equity measure) and labels each institution according to whether it went bankrupt.

## Qualitative Bankruptcy

# Discussion

# References

[1] Maaten, L. v. d.; Hinton, G., Visualizing data using t-SNE. Journal of Machine Learning Research 2008, 9, 2579-2605.

[2] Samuel Kaski, Janne Nikkil¨a, Merja Oja, Jarkko Venna, Petri T¨or¨onen, and Eero Castr´en. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4:48, 2003.