T-SNE Classifier

# Abstract

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

# 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 [1]. Trustworthiness measures the success of the clustering process and is agnostic to labels.

# 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. If the nth bit in the number equals 1 include the nth feature in the dataset.
2. Execute the T-SNE algorithm reducing the dataset to 2 dimensions
3. 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.
4. Repeat quality calculation by using majority vote from 3 neighbors.
5. Repeat quality calculation by using majority vote from 6 neighbors.
6. Calculate trustworthiness.

# 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] 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.