Cluster Weighted Model Based on TSNE Algorithm for High-Dimensional Data

Kehinde Olobatuyi (✉ k.olobatuyi@campus.unimib.it)

University of Milano–Bicocca <https://orcid.org/0000-0002-4635-7895>

Research Article

Keywords: Cluster-weighted model, T-distributed Stochastic Neighbor Embedding, Expectation Maximization algorithm, Parsimonious technique, FlexCWM R package

Posted Date: March 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-347795/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Similar to many Machine Learning models, both accuracy and speed of the Cluster weighted models (CWMs) can be hampered by high-dimensional data, leading to previous works on a parsimonious technique to reduce the effect of "Curse of dimensionality" on mixture models. In this work, we review the background study of the cluster weighted models (CWMs). We further show that parsimonious technique is not sufficient for mixture models to thrive in the presence of huge high-dimensional data. We discuss a heuristic for detecting the hidden components by choosing the initial values of location parameters using the default values in the "FlexCWM" R package. We introduce a dimensionality reduction technique called T-distributed stochastic neighbor embedding (TSNE) to enhance the parsimonious CWMs in high-dimensional space. Originally, CWMs are suited for regression but for classification purposes, all multi-class variables are transformed logarithmically with some noise. The parameters of the model are obtained via expectation maximization algorithm. The effectiveness of the discussed technique is demonstrated using real data sets from different fields.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the latest manuscript can be downloaded and [accessed as a PDF](#).

Figures

Figure 1

Models Used in CWMs clustering: Example of contours of the bivariate normal component densities for the 14 parameterization of the covariance matrix. Source: [3]

Figure 2

The visualization of the descriptive summary of the original Abalone data colored according to the grouping of the Rings: Black (1-8), Red(9-10), and Green (≥ 11)

Figure 3

Model selection for the Abalone data using BIC values of the fourteen models. The BIC produced by three models select the correct number of components.

Figure 4

The classification plot of CWM-tSNE for G =3 with model VVV selected by BIC.

Figure 5

The classification plot of CWM-tSNE for G =4 as suggested by the BIC

Figure 6

Model selection for the protein data using BIC values of the fourteen models. The BIC produced by five models select the correct number of components.

Figure 7

The plot produced by CWMs after dimension reduction via tsne. CWMs selected eight components which aligns to the true class of the localization site of protein.

Figure 8

The tSNE for dimensionality reduction of the Epileptic Seizure data for 1000 iteration, perplexity = 15 and theta = 0.5.

Figure 9

The CWM-tSNE plot for clustering the low-dimensional data produced by tSNE for Seizure recognition data with five categories.

Figure 10

The tSNE for dimensionality reduction of the Epileptic Seizure data for 10,000 iterations, perplexity =250, and theta =0.5.

Figure 11

The CWM-tSNE plot for clustering the low-dimensional representation of Seizure recognition data produced by tSNE with EEE model.

Figure 12

The Model selection of BIC for Seizure data among the fourteen parsimonious model; BIC selected wrong number of mixture component when the true component according to the label is two-categorical.

Figure 13

The Model selection of ICL for Seizure data among the fourteen parsimonious model; ICL selected EEE with the correct number of mixture component when the true component is two categories.