Visualisierung von Musikdaten mittels t-SNE und PCA am Beispiel pgvector

Jannis Gehring
INF22B, Data Warehouse
Duale Hochschule Baden-Württemberg (DHBW)
Stuttgart, Deutschland
inf22115@lehre.dhbw-stuttgart.de

***Abstract*—**

Contents

I. Grundlagen

A. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) ist ein lineares Verfahren zur Dimensionsreduktion. Es wurde in der ersten Hälfte des zwanzigsten Jahrhunderts entwickelt, fand aber aufgrund seiner Berechnungsanforderungen erst in den 60ern breite Anwendung. [1]

Mathematisch liegt PCA eine Eigenvektorberechnung zugrunde. Vereinfacht wird zu Beginn die Kovarianz-Matrix $\boldsymbol{S}$ des Datensatzes berechnet. Dann werden die Eigenvektoren dieser Matrix berechnet und nach dem Betrag ihrer jeweiligen Eigenwerte sortiert. Von diesen n Eigenvektoren werden nun, je nach Anwendungsfall, die k ersten gewählt ($k \leq n$). Mit einer weiteren Matrix $\boldsymbol{W}$, die jene gewählten Eigenvektoren als Zeilenvektoren hat, wird nun die Transformation der Ausgangstupel in den (meist niedriger-dimensionalen) Raum durchgeführt. [2]

Das folgende Beispiel gibt dieser mathematischen Definition eine handfeste Intuition. Hier entspricht PCA dem iterativen Auswählen von zueinander orthogonalen Linien durch den Datensatz, die diesen am besten teilen. *Am besten* bedeutet hier, dass die Varianz der Projektionen auf diese Linie maximal sind. [2]

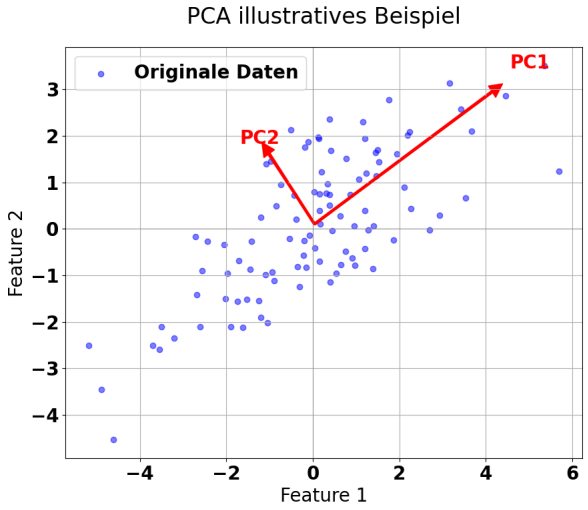

Fig. 1: Veranschaulichung von PCA mit $n = k = 2$

B. t-distributed Stochastic Neighbour Embedding (t-SNE)

t-distributed Stochastic Neighbour Embedding (t-SNE) ist ein nicht-lineares Verfahren zur Dimensionsreduktion. Es wurde 2008 als Weiterentwicklung des Verfahrens Stochastic Neighbour Embedding (SNE) vorgestellt.

Zentral für SNE ist, dass Nachbarschaftsbeziehungen des hochdimensionalen Raums so gut wie möglich im niedrigdimensionalen Raum erhalten bleiben müssen. Hier zu wird im Ausgangsraum X die Wahrscheinlichkeit $p_{j|i}$ definiert. Wenn vom Punkt $x_i \in X$ ein zufälliger Nachbar gewählt wird, wobei nähere Punkte anhand einer Gauß-verteilung wahrscheinlicher sind als fernere, dann bezeichnet $p_{j|i}$ die Wahrscheinlichkeit, dass hierbei der Punkt x_j gewählt wird.

Genau die gleichen Beziehungen werden im niedrigdimensionalen Raum $Y \supset \{y_i, y_j\}$ definiert, wobei die Wahrscheinlichkeit hier mit $q_{j|i}$ bezeichnet wird.

Ziel von SNE ist nun, die Punkte aus X so auf Y abzubilden, dass $q_{j|i}$ für alle i und j möglichst nahe an $p_{j|i}$ ist. Dies wird mittels eines Gradient-Descent Algorithmus durchgeführt.

t-distributed Stochastic Neighbour Embedding (t-SNE) erweitert SNE in zwei Weisen. Zum Einen passt es die Wahrscheinlichkeitsverteilungen so an, sodass $p_{ij} = p_{ji}$ und $p_{ij} = p_{ji}\ \forall i, j$, zum Anderen nutzt es im niedrig-dimension-

alen Raum nicht die Gaußverteilung zur Ermittlung des Nachbarn, sondern die t-Verteilung nach Student.

C. Vergleich von Principal Component Analysis (PCA) und t-SNE

Im Folgenden werden PCA und t-SNE anhand unterschiedlicher Kriterien verglichen [3]:

Kriterium	**PCA**	**t-SNE**
Ziel der Dimensionsreduktion	Maximieren der Varianz	Erhalten von lokaler Struktur
Linearität	Linear	Nicht-Linear
Iterativ?	Ja	Nein
Berechnungskomplexität	$\mathrm{O}(d^2n + n^3)$	$\mathrm{O}(n^2)$

II. Installation

III. Umsetzungsbeispiel

References

[1] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993, doi: https://doi.org/10.1016/0098-3004(93)90090-R.

[2] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation 11(2)*, pp. 443–482, 1999, [Online]. Available: http://www.miketipping.com/papers/met-mppca.pdf

[3] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review*, vol. 40, p. 100378, 2021, doi: https://doi.org/10.1016/j.cosrev.2021.100378.