Příjmení: Vacho	Jméno: Peter	Obor: Softwarové Inženýrství
Předmět: **Softcomputing a Datamining**		Datum: 09.12.2024
Název protokolu: MNIST - redukce dimenzionality		

MNIST - redukce dimenzionality

Úvod

V tejto úlohe sme sa zamerali na aplikáciu redukcie dimenzionality na MNIST dataset pomocou metód PCA a t-SNE. Cieľom bolo zistiť, ako tieto metódy znižujú dimenzionalitu dát a ako ich výsledky ovplyvňujú vizualizáciu a presnosť klasifikácie pomocou KNN.

Princíp PCA a t-SNE

- PCA (Principal Component Analysis) je lineárna metóda redukcie dimenzionality, ktorá transformuje dáta do nového priestoru tak, že maximálne zachová variabilitu v dátach. Tento proces sa dosahuje pomocou vlastných vektorov a vlastných čísel, ktoré reprezentujú hlavne komponenty.
- t-SNE (t-distributed Stochastic Neighbor Embedding) je nelineárna metóda vhodná pre vizualizáciu. Zachováva lokálne vzťahy medzi dátami a je špeciálne navrhnutá na projekciu do 2D alebo 3D priestoru. Používa stochastický algoritmus na minimalizáciu rozdielov medzi distribúciou v pôvodnom a redukovanom priestore.

Implementácia

Implementácia tejto úlohy bola relatívne jednoduchá, pre oba zadané algoritmy som tu využil vstavané funkcie z scikit-learn kni\žnice (sklearn.decomposition.PCA & sklearn.manifold.TSNE).

Najväčším problémom pri implementovaní bolo spracovať všetky dáta, keďže MNIST dataset je relatívne veľký, redukovanie dimenzionality na celom datasete by trvalo dosť dlho. Pôvodný dataset MNIST obsahuje 70 000 riadkov a 784 stĺpcov (28x28 pixelov).Rozhodol som sa tak obmedziť počet riadkov na 8 000 (toto číslo je nastaviteľné). To mi umožnilo signifikantne znížiť dobu trvania programu a dosiahnuté výsledky stále celkom dobre reprezentovali štruktúru MNIST datasetu.

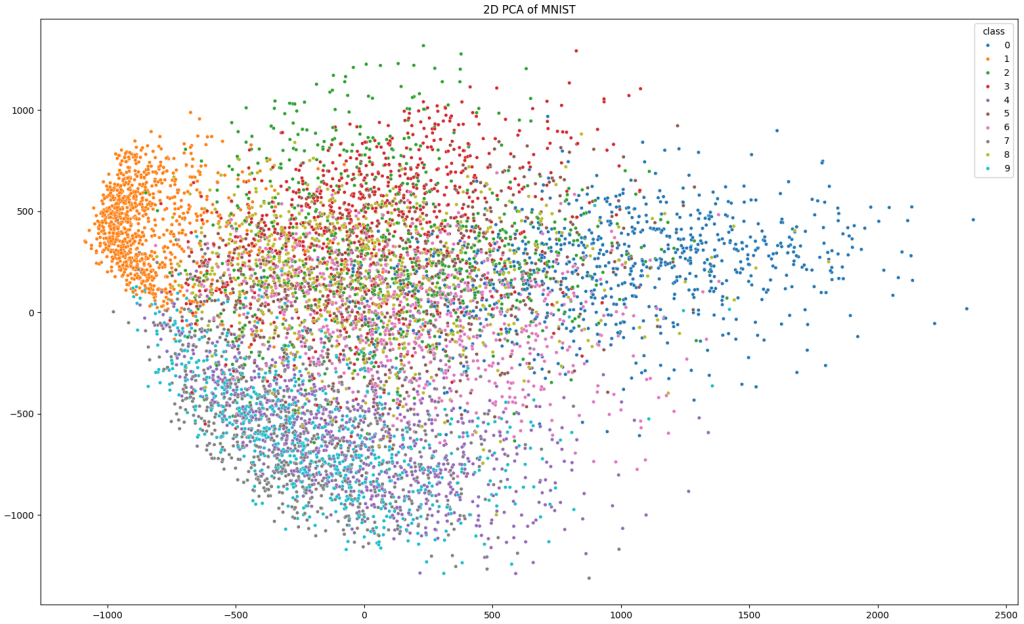
Parametre algoritmov

- Pre PCA som nastavil počet komponentov na 2, aby som mohol vykresliť dáta v dvoj-rozmernom priestore.
- Pri t-SNE som použil nasledujúce parametre:
    - Perplexita: 30
    - Maximálny počet iterácií: 500
    - Počet komponentov: 2 (pre 2D)

Výsledky

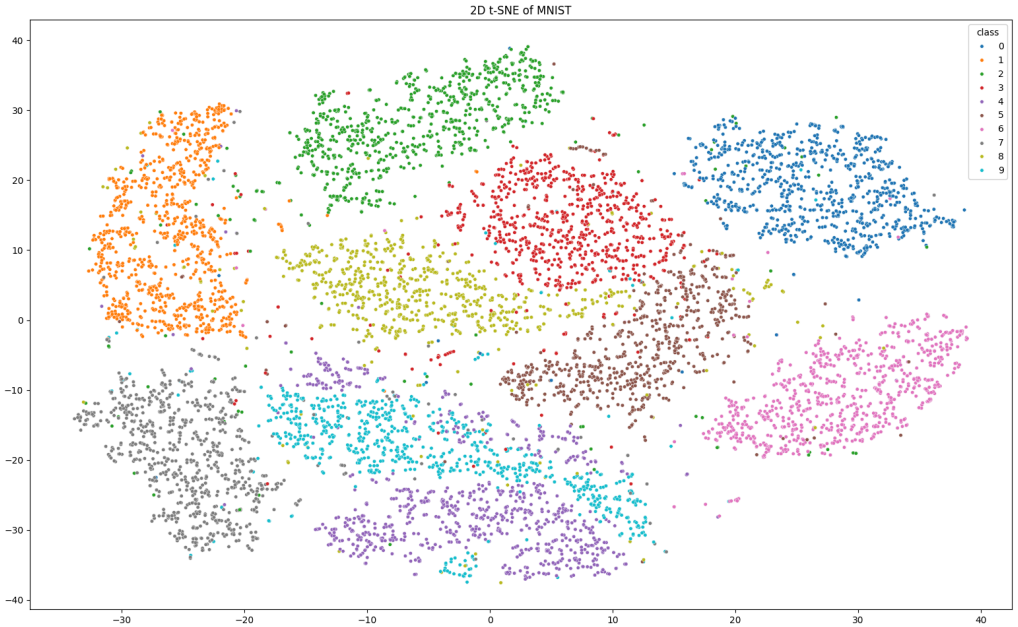
Grafy

Z výsledkov bolo vidieť značný rozdiel medzi týmito algoritmami, kde PCA algoritmus produkoval nádledujúci graf:



Graf redukovaných dát pomocou PCA ukazuje slabšie zoskupenie jednotlivých hodnôt. Tento jav je očakávaný, pretože PCA je lineárna metóda, ktorá nezachytáva nelineárne vzťahy medzi dátami.

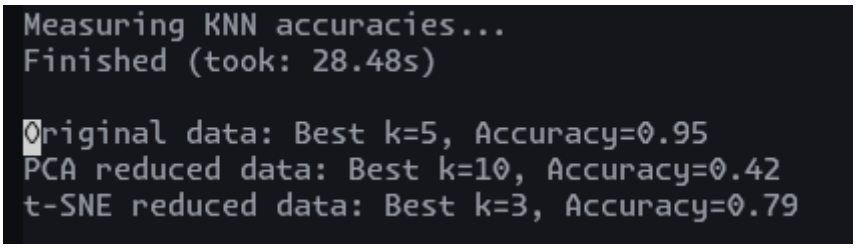
V porovnaní, dáta po redukcií s t-SNE algoritmom vytvorili následujúci graf:



Tu sú jednotlivé clustre dát veľmi dobre oddelené a je vidieť, že táto metóda efektívne zachytáva lokálne konzistencie, čo vedie k lepšiemu rozdeleniu jednotlivých tried.

Presnosti KNN

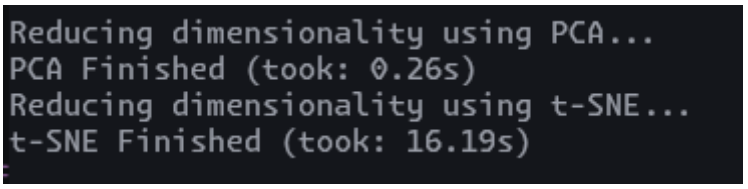
Ďalej som ešte spočítal KNN presnosti pred a po redukcií, kde som dostal nasledujúci výsledok:



- Pre pôvodné dáta bol algoritmus KNN najpresnejší, čo sa dá očakávať, pretože žiadna informácia nebola stratená.
- Po redukcii dimenzionality bola presnosť KNN nižšia, ale t-SNE dosiahlo značne lepšie výsledky ako PCA, pretože lepšie zachytáva nelineárne vzťahy v dátových bodoch.

Rýchlosť algoritmov

Posledným zaujímavím poznatkom bol rozdiel v rýchlosti týchto dvoch metód, zatiaľ čo PCA metóda bol relatívne rýchla t-SNE algoritmus bol časovo značne náročnejší:



Záver

Zadaním tejto úlohy bolo vyskúšať si redukciu dimenzionality datasetu. Redukovanie dimenzionality je užitočná metódla hlavne pre vizualizáciu mnohodimenzionálneho datasetu, keďže ludia nie sú schopný jednoducho spracovať, ba dokonca ani si predstaviť dáta vo viac ako 3 dimenziách- V tomto prípade sme redukovali na 2 dimenzie a zobrazili sme si graf výsledných redukovaných dát pre oba algoritmy.

Z výsledkov sme videli, že PCA je jednoduchá a rýchla metóda, ktorá je vhodná pre lineárne vzťahy v dátach. t-SNE je naopak lepší pre zložitejšie a nelineárne vzťahy a lepšie zobrazuje clustre v dátach.

Redukcia dimenzionality môže byť užitočná v oblastiach ako vizualizácia genomických dát, segmentácia zákazníkov, alebo zjednodušenie predspracovania pre strojové učenie.

Táto úloha bola menej komplexná ako tie predošlé ohľadne implementácie, ale stále bola zaujímavou a poukázala ako funguje redukcia dimenzionality a na čo sa pri analýze dát môže použiť.