

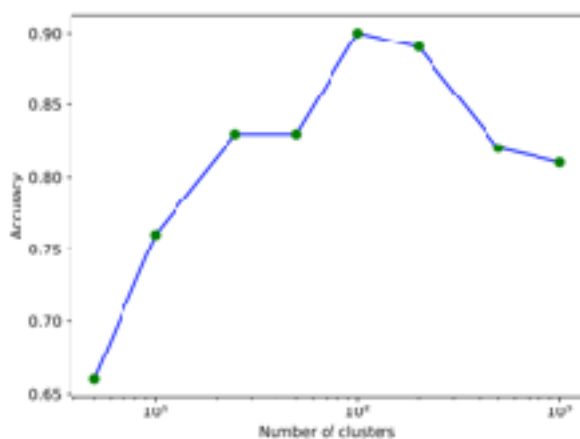
Vedere Artificiala - tema 4

Clasificarea imaginilor folosind modelul Bag-Of-Visual-Words

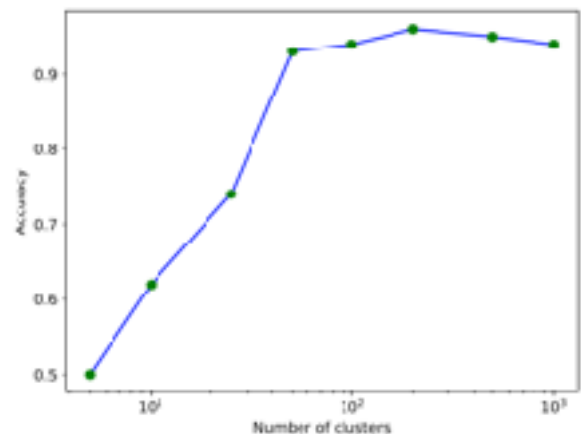
Eric-Petru Stavarache, Lucian Bicsi

Acuratete vs. dimensiunea vocabularului (subpunctele a, b):

Clasificator NN



Clasificator SVM liniar



Cea mai buna acuratete: 0.9 (K = 100)

Cea mai buna acuratete: 0.96 (K = 200)

Evaluarea pe date de antrenare (subpunctul c):

Pentru a nu folosi datele de testare in vederea evaluarii clasificatorilor si pentru a evita problemele grave de overfitting pe care le implica evaluarea modelului pe date care sunt folosite si pentru testare, vom folosi strategii train-validation split:

Vom separa datele (etichetate) de input in doua multimi (cu un raport, sa zicem, de 70/30) in set de antrenare si set de validare. Vom folosi, mai apoi, statisticile obtinute pe setul de validare pentru a decide parametrii optimi. Deoarece modelul a fost tunat pentru a obtine acuratetea optima pe datele de validare, rezultatele sunt o supraestimare.

In final, vom face evaluarea finala pe setul de date de testare. Acuratetea pe care o vom obtine reflecta performanta efectiva a modelului, iar din acest punct nu avem voie sa il mai "tunam" (altfel, riscam supraevaluarea mentionata anterior).

Nedeterminismul de la nivelul antrenarii (subpunctul d):

Ruland algoritmul de clasificare de mai multe ori pentru un vocabular de dimensiune fixata vom obtine rezultate variate (de exemplu, 7 rulari pentru K = 100 si clasificatorul SVM obtin performanta [0.94, 0.92, 0.93, 0.96, 0.92, 0.95, 0.92]). In general, un algoritm este nedeterminist in momentul in care se intampla cazuri de comportament nedefinit (accesari ilegale ale memoriei, apelul gresit al functiilor, etc.) si atunci cand se folosesc **numere generate aleator la executie**.

Empiric, am observat ca aceasta variatie este independenta de clasificatorul ales (NN, SVM), indicand faptul ca sursa nedeterminismului tine de algoritmul de clusterizare (K-Means).

In adevar, algoritmul K-Means porneste de la o generare aleatoare de K centre ale clusterelor, care urmeaza a fi augmentate de setul de date, iar aceasta initializare aleatoare determina alegerea finala a vocabularului. Cum exista o corelatie majora intre centrele vectorilor alesi de algoritmul K-Means ("cuvintele" vocabularului) si performanta clasificatorului, impactul poate fi vazut experimental pe date putine.

O solutie pentru a stabili o evaluare mai buna, este aplicarea unei strategii Monte-Carlo (informal, vom rula algoritmul pentru un K fixat de mai multe ori, iar performanta va fi caracterizata de acuratetea medie in cadrul rularilor — media aritmetica).