Determinarea tipului de erorare a plăcilor de oțel utilizând algoritmul de „Random forest”

Răşcanu Rareș-Augustin (**421A**) | ISIA | 15.11.2022

Codul sursă

Este valabil pe GitHub la următoarea adresă:
https://github.com/Disreality/ETTI-ISIA-Project

1. Problema

Trebuie determinat tipul de eroare al fiecărei dintre cele **1941 exemple** de plăci de oțel dintr-un anumit set. În cele **34 de dimensiuni** ale bazei de date, **27** dintre acestea se referă la diferite proprietăți ale materialului (perimetru, vârfuri etc.), iar **7** descriu, binar, tipul erorii determinat în mod experimental. Mai precis, aceste ultime dimensiuni se aseamănă unui vector binar ce are o singură valoare maximă la indicele tipului de eroare dat spre semnalare.

2. Librării & considerente tehnice

Librăria ce a permis analizarea datelor este **scikit-learn**. Pentru preluarea datelor din baza de date în Python s-au folosit unelte din modulul **csv**. Diferite variabile globale, funcții de „back-end" și „front-end" au fost declarate & definite în prima secțiune a codului sursă. Funcția **getType**, responsabilă pentru a determina tipul de eroare a plăcii pe baza informațiilor furnizate din baza de date, conține diferitele denumiri ale acestora („Pastry", „Bumps" etc.).

3. Seturi de date (train & test)

Baza de date a fost folosită în întregime în ambele etape, diferența constând în faptul că informațiile referitoare la tipul erorii au fost considerate doar în prima etapă. Am putut astfel să calculăm procentajul de răspunsuri corecte folosind următoarea formulă, derivată din formula procentajului:

$$procentaj\ răspunsuri\ corecte = 100 * \frac{număr\ discret\ de\ răspunsuri\ corecte\ din\ test}{număr\ total\ de\ exemple\ din\ baza\ de\ date}$$

4. Rezultate (metrică & variații)

S-au construit toate combinațiile existente pe baza valorilor pentru procentajul „in-bag" și al numărului de dimensiuni menționate în documentul ce include descrierea proiectului (25%, 50%, 85%, respectiv 10%, 50%, 80%). Metrica folosită este cea calculată pe baza formulei menționate la punctul anterior. Aceste informații sunt afișate, de asemenea, în momentul rulării codului sursă.

Se observă ca rata de succes crește concomitent cu creșterea valorilor pentru procentajul „in-bag" și a numărului de dimensiuni valabile. Acest fapt nu este suprinzător, dat fiind că odată cu creșterea acestor valori crește și numărul de informații valabile algoritmului la momentul formării unui răspuns.