Petrescu Ștefan-Adrian

424B

Proiect ISIA
-raport-

Proiectul se rezolvă cu ajutorul sistemului de clasificare Random Forest. Tema acestuia este Fertility.

În această bază de date 100 de voluntari au donat monstre de spermatozoizi care au fost analizate. Concentrația acestora are legătură cu mediul social, factori de mediu, hobby-uri și probleme medicale. Pe baza acestei baze de date se dorește prezicerea nivelului de concentrație (Normal-N, Alterat-O). Fiecare voluntar prezintă 9 trăsături (dimensiuni) și rezultatul analizei. În situația prezentată, se rezolvă o problemă de clasificare.

Librăria de bază utilizată este sklearn, una din cele mai utile librării pentru machine learning din Python. Din aceasta a fost utilizată metoda ensemble, ce combină mai mulți estimatori de bază cu un anume algoritm pentru a obține un rezultat mai bun (Random Forest utilizează mai mulți arbori). Pe lângă aceasta, a fost utilizată și metoda metrics pentru măsurarea preciziei.

Împărțirea între „train” și „test” a fost realizată în procentul 75-25. Baza de date conținând 100 de eșantioane, primele 75 reprezintă datele de „train”, iar ultimele 25 datele de „test”. În aceeași măsură au fost împărțite și etichetele corespunzătoare fiecărui eșantion : 75 etichete de „train” și 25 etichete de „test”. Datele sunt reprezentate de 2 matrice, iar etichetele de 2 vectori.

Metrica folosită pentru măsurarea performanței (preciziei) este accuracy_score(y_true, y_pred) ce primește ca parametrii etichetele de „test” (y_true) și predicția făcută pe datele de „test” (y_pred). Aceasta compară etichetele adevărate cu etichetele prezise și determină acuratețea cu care sistemul de clasificare a lucrat.

În proiect s-a utilizat un Random Forest Classifier (trăsăturile pot lua doar anumite valori) cu 10 arbori (n_estimators = 10) care împarte cele 75 de eșantioane în cei 10 arbori în ordine aleatoare. Se folosește principiul de bagging, prin intermediul căruia se specifică, printr-un parametru, numărul maxim de date aleatoare care este utilizat în fiecare arbore. Acest parametru este variat prin procentul „in-bag” (max_samples). Celălalt parametru, care se variază, este numărul maxim de trăsături disponibile pentru utilizare în fiecare nod al arborelui. Acesta reprezintă numărul de dimensiuni alese într-un nod (max_features). Folosirea simultană a principiului de bagging și de subspații de trăsături duce la diversitate mare a arborilor ce compun un Random Forest, asigurând performanță sporită.

Procentul „in-bag”(%)	Numărul de dimensiuni(%)	Acuratețea(%)
25	10	92
50	10	92
85	10	88
25	50	88
50	50	92
85	50	92
25	80	88
50	80	92
85	80	96

Acuratețea nu variază mult între schimbările parametrilor. Acest lucru se datorează setului de date dezechilibrat. Din totalul de 100 de voluntari, 88 au eticheta Normal-N, în timp ce doar 12 au eticheta Alterat-O. Așadar, setul de antrenare va conține puține diagnostice de tip O și multe de tip N. Acest lucru se aplică și în cazul celui de testare, rezultând o foarte bună recunoaștere a tipului N și o recunoaștere mai slabă a tipului O.