Clasificare bazată pe ansambluri. Random Forest

Random Forest este o metodă de învățare bazată pe ansamburi ce operează cu noțiunea de arbore de decizie, studiată în laboratorul anterior. Metodele bazate pe ansambluri funcționează pe principiul conform căruia un grup de „clasificatori slabi" (*weak classifiers*) pot forma, împreună, un clasificator puternic (*strong classifier*). Astfel, rezultatul clasificării folosind un ansamblu se determină prin compunerea rezultatelor individuale ale mai multor modele de clasificare. Principiul se poate enunța astfel: decizia unei populații de indivizi este mai fiabilă decât decizia unui singur individ, atât timp cât deciziile individuale ale membrilor populației sunt de o calitate decentă.

Un arbore de decizie singular se construiește pornind de la întregul set de date de antrenare, conform cu cele studiate în laboratorul anterior. Un astfel de arbore este puternic influențat de modificări ale datelor de antrenare: orice schimbare a datelor poate conduce la modificări semnificative ale structurii arborelui. De asemenea, un arbore singular este susceptibil la fenomenul de *overfitting* – arborele are eroare de clasificare redusă pe setul de date de antrenare, dar nu generalizează suficient de bine – eroare mare de clasificare pentru alte date de test.

În cazul metodei Random Forest, se folosește o multitudine de arbori de decizie, astfel:

- fiecare arbore se generează pornind de la o submulțime a datelor de antrenare inițiale și/sau folosind o submulțime a atributelor datelor (Fig. 1)
- submulțimile de date și atribute se generează aleatoriu (instanțele și atributele se pot repeta de la o submulțime la alta)
- o instanță se clasifică separat de către fiecare arbore. Rezultă câte o decizie (câte o clasă) pentru fiecare arbore
- clasa finală este cea majoritară (clasa care apare cel mai frecvent printre mulțimea de clase furnizată de arbori)

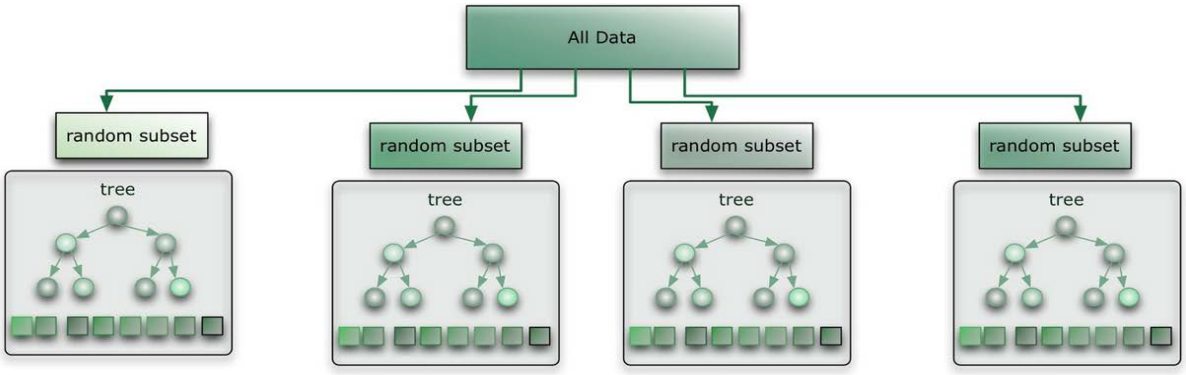

Fig.1. Generarea de arbori în cadrul algoritmului Random Forest

În cele ce urmează vom ilustra modul de generare și utilitatea unui random forest printr-un exemplu: pornim cu un set de date de antrenare (Tabelul 1) și unul de test (Tabelul 2). Vom construi un arbore singular și un random forest pornind de la aceleași date antrenare. Apoi, folosind setul de date de test, vom determina și vom compara erorile de clasificare ale arborelui singular cu cele ale random forest.

Tabelul 1. Setul de date de antrenare

Starea vremii	Temperatura	Umiditate	Vant	Joc
Ploaie	Mare	Mare	Absent	Nu
Ploaie	Mare	Mare	Prezent	Nu
Innorat	Mare	Mare	Absent	Da
Soare	Medie	Mare	Absent	Da
Soare	Mica	Normala	Absent	Da
Soare	Mica	Normala	Prezent	Nu
Innorat	Mica	Normala	Prezent	Da
Ploaie	Medie	Mare	Absent	Nu
Ploaie	Mica	Normala	Absent	Da
Soare	Medie	Normala	Absent	Da
Ploaie	Medie	Normala	Prezent	Da
Innorat	Medie	Mare	Prezent	Da
Innorat	Mare	Normala	Absent	Da
Soare	Medie	Mare	Prezent	Nu

Tabelul 2. Setul de date de test

Starea vremii	Temperatura	Umiditate	Vant	Joc
Ploaie	Mare	Mare	Absent	Da
Ploaie	Mare	Mare	Prezent	Nu
Innorat	Medie	Normala	Absent	Da
Innorat	Medie	Mare	Absent	Nu
Soare	Mica	Normala	Absent	Da
Soare	Mica	Normala	Prezent	Da
Ploaie	Mica	Normala	Prezent	Da
Ploaie	Medie	Mare	Absent	Nu
Ploaie	Mica	Normala	Absent	Da
Soare	Medie	Normala	Absent	Da
Ploaie	Mare	Normala	Prezent	Nu
Innorat	Mare	Mare	Prezent	Nu
Innorat	Mare	Normala	Absent	Da
Soare	Medie	Mare	Absent	Nu

Arborele singular:

Pentru obținerea arborelui, vom proceda asemănător cu generarea arborelui ID3 (laboratorul anterior), însă pentru a sorta atributele funcție de importanța lor decizională vom folosi indicele Gini, calculat conform cu Ecuația 1.

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \quad (1)$$

Unde:

S este setul de date pentru care se calculează indicele Gini

C este numărul de clase (în setul de date din tabelul anterior sunt două clase, așadar, C = 2)

p_i este probabilitatea de apariție a clasei *i* în setul de date S

În cazul setului de date anterior,

$Gini(S) = 1 – (p(Da)^2 + p(Nu)^2)$

$p(Da) = 9/14$, $p(Nu) = 5/14$, așadar

$Gini(S) = 1 – (9/14)^2 – (5/14)^2 = 0.459$

Ca și în cazul arborilor ID3, pentru arborele întreg și pentru fiecare subarbore al său, trebuie să determinăm atributul care are cea mai mare importanță decizională – *atributul cu indicele Gini minim*. Calculul indicilor Gini ai atributelor se realizează pe același principiu ca și determinarea entropiei. Un atribut A cu valorile V1, V2, V3 cauzează partiționarea setului de date S în trei submulțimi SV1, SV2, SV3, fiecare corespunzând unei valori a atributului. SV1 este submulțimea lui S care conține doar instanțele cu valoarea V1 a atributului A, analog pentru SV2, SV3. În acest caz, indicele Gini care rezultă în urma partiționării datelor folosind atributul A se determină astfel:

$Gini(S,A) = p(V1) * Gini(SV1) + p(V2) * Gini(SV2) + p(V3) * Gini(SV3)$

Unde:

p(V1) este probabilitatea ca atributul A să aibă valoarea V1

Gini(SV1) este indicele Gini determinat pentru submulțimea SV1

Idem pentru V2 și V3

Ca valoare a unui nod al arborelui, se utilizează atributul cu indicele Gini minim, iar partiționarea ulterioară a datelor se realizează folosind valorile acelui atribut.

Calculele sunt foarte similare cu cele efectuate în cazul arborilor ID3 din laboratorul anterior, diferența esențială este utilizarea indicelui Gini în locul entropiei.

Pentru datele din Tabelul 1 rezultă următorul arbore:

```
Starea vremii
              --Innorat--> Da
              --Soare--> Vant
                            --Prezent--> Nu
                            --Absent--> Da
              --Ploaie--> Umiditate
                                    --Normala--> Da
                                    --Mare--> Nu
```

Determinăm eroarea de clasificare a acestui arbore folosind datele de test: în Tabelul 3 se observă că acest arbore nu reușește să clasifice corect anumite instanțe. Din cele 14 instanțe 6 sunt clasificate greșit, așadar **eroarea de clasificarea a arborelui este 6/14 = 0.428 .**

Tabelul 3. Setul de date de test: coloana ST conține clasele prezise de arborele singular, iar coloana Joc conține clasele din setul de date (cele "corecte")

Starea vremii	Temperatura	Umiditate	Vant	ST	Joc
Ploaie	Mare	Mare	Absent	Nu	Da
Ploaie	Mare	Mare	Prezent	Nu	Nu
Innorat	Medie	Normala	Absent	Da	Da
Innorat	Medie	Mare	Absent	Da	Nu
Soare	Mica	Normala	Absent	Da	Da
Soare	Mica	Normala	Prezent	Nu	Da
Ploaie	Mica	Normala	Prezent	Da	Da
Ploaie	Medie	Mare	Absent	Nu	Nu
Ploaie	Mica	Normala	Absent	Da	Da
Soare	Medie	Normala	Absent	Da	Da
Ploaie	Mare	Normala	Prezent	Da	Nu
Innorat	Mare	Mare	Prezent	Da	Nu
Innorat	Mare	Normala	Absent	Da	Da
Soare	Medie	Mare	Absent	Da	Nu

Random forest:

Presupunem că "pădurea" este formată din n arbori:

- pornind de la setul de date de antrenare (Tabelul 1), generăm aleatoriu
  - o n submulțimi ale setului de date
  - o n submulțimi ale listei atributelor

- pentru fiecare submulțime a setului de date și submulțime de atribute vom construi un arbore separat, prin aceeași metodă care s-a utilizat la generarea arborelui singular menționat anterior

Exemplu: 5 arbori generați pornind de la 5 submulțimi aleatorii ale setului de date de antrenare și 5 liste ce conțin o parte din atributele datelor.

Mai jos, pentru fiecare arbore Tree 1-5 se prezintă, în ordine:

- structura arborelui
- atributele care s-au luat în calcul la generarea arborelui
- submulțimea datelor de antrenare pe baza căreia s-a generat arborele

```
Tree 1:
 Starea vremii
              --Soare--> Temperatura
                                 --Mare--> Da
                                 --Medie--> Da
                                 --Mica--> Da
              --Ploaie--> Temperatura
                                  --Mare--> Nu
                                  --Medie--> Nu
                                  --Mica--> Da
              --Innorat--> Da
```

```
['Temperatura', 'Starea vremii']
```

	Starea vremii	Temperatura	Umiditate	Vant	Joc
4	Soare	Mica	Normala	Absent	Da
2	Innorat	Mare	Mare	Absent	Da
1	Ploaie	Mare	Mare	Prezent	Nu
8	Ploaie	Mica	Normala	Absent	Da
12	Innorat	Mare	Normala	Absent	Da
0	Ploaie	Mare	Mare	Absent	Nu
5	Soare	Mica	Normala	Prezent	Nu
7	Ploaie	Medie	Mare	Absent	Nu
11	Innorat	Medie	Mare	Prezent	Da
9	Soare	Medie	Normala	Absent	Da
6	Innorat	Mica	Normala	Prezent	Da

```
Tree 2:
 Vant
     --Absent--> Da
     --Prezent--> Temperatura
                              --Mare--> Nu
                              --Medie--> Da
                              --Mica--> Da
```

```
['Vant', 'Temperatura', 'Starea vremii']
```

	Starea vremii	Temperatura	Umiditate	Vant	Joc
12	Innorat	Mare	Normala	Absent	Da
2	Innorat	Mare	Mare	Absent	Da
10	Ploaie	Medie	Normala	Prezent	Da
8	Ploaie	Mica	Normala	Absent	Da
11	Innorat	Medie	Mare	Prezent	Da
1	Ploaie	Mare	Mare	Prezent	Nu

```
Tree 3:
 Starea vremii
              --Soare--> Vant
                               --Absent--> Da
                               --Prezent--> Nu
              --Ploaie--> Temperatura
                                  --Mare--> Da
                                  --Medie--> Nu
                                  --Mica--> Da
              --Innorat--> Da
```

```
['Starea vremii', 'Vant', 'Temperatura']

   Starea vremii Temperatura Umiditate     Vant Joc
3          Soare       Medie      Mare   Absent  Da
7         Ploaie       Medie      Mare   Absent  Nu
11       Innorat       Medie      Mare  Prezent  Da
8         Ploaie        Mica   Normala   Absent  Da
5          Soare        Mica   Normala  Prezent  Nu



Tree 4:
 Starea vremii
              --Soare--> Vant
                             --Absent--> Da
                             --Prezent--> Nu
              --Ploaie--> Vant
                              --Absent--> Nu
                              --Prezent--> Nu
              --Innorat--> Da


['Starea vremii', 'Vant']

   Starea vremii Temperatura Umiditate     Vant Joc
0         Ploaie        Mare      Mare   Absent  Nu
3          Soare       Medie      Mare   Absent  Da
11       Innorat       Medie      Mare  Prezent  Da
12       Innorat        Mare   Normala   Absent  Da
8         Ploaie        Mica   Normala   Absent  Da
13         Soare       Medie      Mare  Prezent  Nu
9          Soare       Medie   Normala   Absent  Da
7         Ploaie       Medie      Mare   Absent  Nu
5          Soare        Mica   Normala  Prezent  Nu
4          Soare        Mica   Normala   Absent  Da
2        Innorat        Mare      Mare   Absent  Da



Tree 5:
 Vant
     --Absent--> Da
     --Prezent--> Temperatura
                             --Mare--> Nu
                             --Medie--> Nu
                             --Mica--> Umiditate
                                              --Mare--> Da
                                              --Normala--> Da


['Temperatura', 'Umiditate', 'Vant']

   Starea vremii Temperatura Umiditate     Vant Joc
5          Soare        Mica   Normala  Prezent  Nu
2        Innorat        Mare      Mare   Absent  Da
13         Soare       Medie      Mare  Prezent  Nu
6        Innorat        Mica   Normala  Prezent  Da
4          Soare        Mica   Normala   Absent  Da
8         Ploaie        Mica   Normala   Absent  Da
```

Clasificăm datele din setul de test folosind cei 5 arbori, individual. Clasa determinată de întreaga “pădure” **va fi cea obținută de majoritatea arborilor**. Comparăm apoi eroarea care rezultă în urma acestei clasificări cu cea determinată pentru arborele singular. Rezultatele sunt prezentate în Tabelul 4. Se observă că Random Forest reușește să clasifice corect mai multe instanțe decât arborele singular.

Random Forest a clasificat greșit 3 instanțe din cele 14, așadar **eroarea de clasificare este 3/14 = 0.214**. Amintim faptul **că arborele singular avea o eroare de 0.428 .**

Tabelul 4. Evaluarea clasificării folosind Random Forest și un arbore singular. Coloanele T1-5 conțin clasele determinate folosind cei 5 arbori din Random Forest. Coloana RF conține decizia majoritară a celor 5 arbori. Coloana ST conține clasele determinate de arborele singular (aceleași ca în Tabelul 3). Coloana Joc conține clasele din setul de date de test (cele "corecte")

Starea vremii	Temperatura	Umiditate	Vant	T1	T2	T3	T4	T5	RF	ST	Joc
Ploaie	Mare	Mare	Absent	Nu	Da	Da	Nu	Da	**Da**	**Nu**	**Da**
Ploaie	Mare	Mare	Prezent	Nu	Nu	Da	Nu	Nu	**Nu**	**Nu**	**Nu**
Innorat	Medie	Normala	Absent	Da	Da	Da	Da	Da	**Da**	**Da**	**Da**
Innorat	Medie	Mare	Absent	Da	Da	Da	Da	Da	**Da**	**Da**	**Nu**
Soare	Mica	Normala	Absent	Da	Da	Da	Da	Da	**Da**	**Da**	**Da**
Soare	Mica	Normala	Prezent	Da	Da	Nu	Nu	Da	**Da**	**Nu**	**Da**
Ploaie	Mica	Normala	Prezent	Da	Da	Da	Nu	Da	**Da**	**Da**	**Da**
Ploaie	Medie	Mare	Absent	Nu	Da	Nu	Nu	Da	**Nu**	**Nu**	**Nu**
Ploaie	Mica	Normala	Absent	Da	Da	Da	Nu	Da	**Da**	**Da**	**Da**
Soare	Medie	Normala	Absent	Da	Da	Da	Da	Da	**Da**	**Da**	**Da**
Ploaie	Mare	Normala	Prezent	Nu	Nu	Da	Nu	Nu	**Nu**	**Da**	**Nu**
Innorat	Mare	Mare	Prezent	Da	Nu	Da	Da	Nu	**Da**	**Da**	**Nu**
Innorat	Mare	Normala	Absent	Da	Da	Da	Da	Da	**Da**	**Da**	**Da**
Soare	Medie	Mare	Absent	Da	Da	Da	Da	Da	**Da**	**Da**	**Nu**

Cerințe:

1) Implementați metoda de generare a unui arbore pornind de la un set de date, folosind indicele Gini pentru selecția atributelor. Generați un arbore folosind setul de date data_vreme3.csv și determinați eroarea sa de clasificare folosind datele de test din data_vreme5.csv

2) Generați un număr oarecare n de submulțimi aleatorii ale datelor in data_vreme3. Pentru fiecare submulțime, alegeți aleatoriu 2 sau 3 atribute dintre cele 4 ale datelor. Generați câte un arbore pentru fiecare submulțime și listă de atribute. Determinați eroarea de clasificare a "pădurii" formate din cei n arbori, așa cum s-a demonstrat în documentația de la laborator. Comparați eroarea cu cea a arborelui determinat la 1).