Raport

Modele și metode validate: rețele Bayesiene, reguli de asociere și metode de clasificare multi-etichetă

efectuat în cadrul proiectului *Abordarea bioeconomică a agenților antimicrobieni – utilizare și rezistență*

(cod - PN-III-P1-1.2-PCCDI-2017-0361).

Colectiv de redacție:

Raluca Mureşan, Claudia Zaharia: reţele Bayesiene şi reguli de asociere în analiza rezistenţei la antibiotice (sect. 2)

Kristian Miok, Daniela Zaharie: clasificare multi-etichetă în analiza rezistenţei la antibiotice (sect. 1, sect. 3, sect. 4)

Coordonator: Daniela Zaharie

Membri:Claudia Zaharia, Raluca Mureşan, Kristian Miok, Radu Moleriu

Data finalizării: 20.11.2020

Raport: R.5.1.3.

Versiunea: 1.0

Acknowledgements

Activities under this work were carried out in the *West University of Timisoara* - financed by project "A bio-economical approach of the antimicrobial agents - use and resistance", in the frame of contract PCCDI 7/19.03.2018, code: PN-III P1-1.2-FPRD2017.

1. Introducere

Analiza datelor corelate cu rezistenţa multiplă la substanţe antimicrobiene poate fi abordată din diferite perspective în funcţie de tipul de date disponibile. În cazul în care se utilizează doar date fenotipice care asociază fiecărei probe o categorie (rezistent, susceptibil, ...) atunci se pot analiza asocierile dintre antibioticele pentru care există rezistenţă cu scopul de a identifica eventuale tipare de rezistenţă. În cazul în care disponibile şi date genotipice se pot construi modele de predicţie care să exploateze corelaţiile dintre mutaţiile observate în anumite gene şi prezenţa rezistenţei la anumite antibiotice. În acest context prezentul raport include analize şi rezultate corespunzătoare celor două direcţii:

- Utilizarea informaţiilor extrase prin analiza asocierilor (folosind metode identificate în etapa anterioară a proiectului) pentru construirea unor reţele Baysiene aditive care permit identificarea unor tipare de asocieri între antibiotice. Rezultatele obţinute au fost prezentate la EHB2020 – E-Health and Bioengineering Conference iar lucrarea (Mureşan et al., 2020) a fost acceptată pentru publicare în volumul conferinţei editat de IEEE.
- Utilizarea unor metode de clasificare multi-etichetă în predicţia antibioticelor la care este indusă rezistenţa prin prezenţa de mutaţii în cadrul unor gene. Au fos analizate mai multe variante de metode şi rezultate comparative preliminare au fost prezentate în (Vladu & Zaharie, 2020).

2. Reţele Bayesiene şi reguli de asociere în analiza rezistenţei la antibiotice

2.1. Reţele Bayesiene

2.1.1. Introducere şi definiţii

Reţelele bayesiene aditive (ABN) reprezintă o tehnică de analiză statistică a datelor ce vizează determinarea un model care să descrie structura de probabilitate şi codependenţele dintr-un set de date. Este o metodă multivariată în care toate variabilele sunt considerate posibili predictori şi oricare dintre variabile pot fi predicţionate. Modelul ABN poate fi utilizat pentru seturi de date complexe şi corelate. Este în general o tehnică de modelare nesupervizată care nu necesită cunoştinţe specifice domeniului studiat, dar există şi posibilitatea ca acestea să fie încorporate în model, transformându-l într-unul semi-supervizat.
ABN poate fi privit ca un model multidimensional de regresie, un analog direct al modelului liniar generalizat (GLM) în care toate variabilele pot fi considerate dependente. Sunt analizate relaţiile dintre variabilele modelului, diferenţiind cele doar corelate de cele ce au

o dependenţă directă. Această metodă se diferenţiază de alte analize statistice (modele liniare generalizate, modele liniare generalizate mixte), care pleacă de la premiza că variabilele covariate sunt independente. Această tehnică a fost folosită cu succes în domenii precum medicină, biologie, ecologie şi epidemiologie veterinară, unde studiile au de regulă date observaţionale şi este necesară modelarea unei *reţele complexe de interacţiuni*.

O reţea bayesiană aditivă este un graf orientat aciclic (directed acyclic graph - DAG) în care fiecare nod reprezintă o variabilă a modelului şi arcele descriu dependenţele dintre acestea. Graful orientat aciclic este reprezentarea grafică a distribuţiei de probabilitate comune tuturor variabilelor din model, parametrii acestuia reprezentând distribuţiile de probabilitate locale ale variabilelor.

Un model ABN este un caz particular de reţea bayesiană $\mathcal{B}$, care pentru un set de variabile $X = \{X_1, X_2, \dots, X_N\}$ constă din:

- graful orientat aciclic $S = (V, E)$, unde:
  - V este o mulţime finită de vârfuri sau noduri şi E este o mulţime finită de arce orientate între noduri;
  - nodurile nu formează cicluri orientate;
  - unui nod $j \in \{1,2, \dots, N\}$îi corespunde o variabilăX_j;
  - mulţimea părinţilor unui nod j este notată Pa_j; un nod j se numeşte părintele nodului k dacă există un arc orientat de la j la k; nodul k se numeşte copilul nodului j;
- mulţimea distribuţiilor de probabilitate locale $\theta_{\mathcal{B}}$; fiecare nod j, având mulţimea părinţilor Pa_j, este parametrizat de o distribuţie de probabilitate locală $P(X_j|Pa_j)$.

Un model ABN se defineşte prin intermediul perechii$\mathcal{A} = (S, \beta_{\mathcal{A}})$, unde S este structura grafului orientat, iar $\beta_{\mathcal{A}}$ este setul de parametri ai modelului. Astfel, pentru a determina modelul, trebuie să specificăm structura grafului orientat şi o mulţime de distribuţii de probabilitate locale. Fiecare dintre acestea este modelată cu ajutorul unei regresii multinomiale sau binomiale logistice, în care X_j este variabila dependentă, iar predictorii sunt variabilele din mulţimea Pa_j.

Nodurile reţelei aditive bayesiene prezintă atât dependenţe condiţionale, cât şi marginale. Principalul obiectiv al modelului este de a exprima relaţiile de independenţă condiţională a variabilelor prin separare grafică, adică de a factoriza distribuţia de probabilitate globală astfel:

$$P(X) = \prod_{j=1}^{N} P(X_j|Pa_j)$$

Pentru a determina aceste distribuţii locale de probabilitate, se consideră S_j numărul de valori ale variabilei X_j şi $s \in \{1,2, \dots, S_j\}$. De asemenea fie $C_j = \prod_{p,X_p \in Pa_j} S_p$ numărul de configuraţii ale părinţilor variabilei X_jşi c o astfel de configuraţie. Probabilitatea $P(X_j = s|Pa_j = c) = \theta_{jcs}$ reprezintă probabilitatea ca $X_j = s$, ştiind că configuraţia pentru mulţimea Pa_jeste c. Au loc următoarele relaţii:

Contract 7PCCDI, COD: PN-III-P1-1.2-PCCDI-2017-0361

$$\theta_{jc} = \bigcup_{s=1}^{S_j} \{\theta_{jcs}\}$$

$$\theta_j = \bigcup_{c=1}^{C_j} \{\theta_{jc}\}$$

$$\theta_{\mathcal{B}} = \bigcup_{j=1}^{N} \{\theta_j\}$$

Folosind această parametrizare, distribuţia comună de probabilitate se factorizează astfel:

$$P(X|\theta_{\mathcal{B}}, S) = \prod_{j=1}^{N} P\left(X_j \middle| Pa_j = c, \theta_{jc}\right)$$

În continuare vor fi prezentate reţele aditive bayesiene în care variabilele au o distribuţie bernoulliană, adică au doar două valori. În acest caz, probabilităţile condiţionate $P\left(X_j = 1 \middle| Pa_j\right) = \theta_{jc1}$ se modelează cu ajutorul regresiei logistice binomiale:

$$\theta_{jc1} = \frac{e^{\beta_{jc1}}}{1 + e^{\beta_{jc1}}}, deci\ \beta_{jc1} = \log\left(\frac{\theta_{jc1}}{1 - \theta_{jc1}}\right)$$

Relaţia de mai sus arată legătura dintre parametrii reţelei bayesiene aditive $\beta_{\mathcal{A}}$ şi parametrii reţelei bayesiene uzuale $\theta_{\mathcal{B}}$. Noutatea reţelelor bayesiene aditive este că parametrii (aditivi) ai modelului β_{jcs} nu sunt modelaţi cu ajutorul unor tabele de contingenţă (ca parametrii θ_{jcs}ai reţelelor bayesiene uzuale), ci cu ajutorul unor modele de regresie logistică.

2.1.2. Determinarea modelului ABN

Fie $\mathcal{D}$ un set de date. Atunci are loc:

$$P(\mathcal{A}|\mathcal{D}) = P(\beta_{\mathcal{A}}, S|\ \mathcal{D}) = P(\beta_{\mathcal{A}}|\ S, \mathcal{D}) \cdot P(S|\mathcal{D})$$

Determinarea structurii şi a parametrilor sunt interconectate, dependente şi necesare pentru determinarea modelului final [Jensen, 2001].

Procesul presupune calcularea pentru fiecare reţea candidată a unui anumit scor şi alegerea reţelei cu scorul cel mai mare. Folosind această metodă se caută reţeaua care reprezintă cel mai bine datele, dar nu este prea complexă. Scorul fiecărei reţele reflectă cât de probabil este ca datele să fi fost generate folosind reţeaua respectivă. Astfel, problema determinării structurii reţelei este una de optimizare.

Structura poate fi deci determinată construind toate DAG-urile posibile şi selectându-l pe acela cu scorul cel mai mare. Astfel este necesar a se preciza o *metodă de căutare* si o *metrică (funcţie scor)* care trebuie să menţină un echilibru între acurateţea şi complexitatea

structurii şi să poate fi calculată cu uşurinţă. În plus este de dorit ca aceasta să fie decompozabilă, adică să aibă loc:

$$scor(\mathcal{D}, \mathcal{S}) = \sum_{j=1}^{N} scor(X_j, Pa_j, \mathcal{D})$$

Un exemplu de metrică bună, care conţine atât un termen ce măsoară cât de bine modelul aproximează datele, cât şi unul care controlează şi penalizează complexitatea acestuia este Bayesian Information Criterion (BIC) [Bernardo, Smith, 2000].

O alternativă la acesta o reprezintă scorul de verosimilitate marginală [MacKay, 2003], metoda bayesiană clasică de a măsura cât de bună este o structură candidată S. Mai exact are loc:

$$P(S|\mathcal{D}) = \frac{P(\mathcal{D}|S)P(S)}{P(\mathcal{D})}$$

unde $P(S)$ este distribuţia apriori a structurii şi $P(\mathcal{D}|S)$ este verosimilitatea marginală a structurii condiţionată de setul de date. Distribuţia apriori a structurii se alege astfel încât să fie uşor de calculat (de regulă se presupune că toate structurile sunt echiprobabile). Scorul BIC al unui model este o aproximare asimptotică a verosimilităţii marginale a modelului respectiv, şi este echivalent cu scorul "minimum description length" introdus în [Rissanen, 1987].

Găsirea structurii optime poate fi făcută printr-o căutare exactă (se determină un model optim global) sau euristică, atunci când numărul variabilelor este mare (modelul determinat este doar optim local). Dacă numărul de variabile este sub 20, se poate face o căutare exactă pe mulţimea ordinilor asociate structurilor [Friedman, Koller 2003], folosind Transformarea Möbius Rapidă introdusă de Koivisto şi Sood în [Koivisto, Sood 2004]. Odată determinată structura modelului ABN, parametrii acestuia sunt determinaţi folosind principiul verosimilităţiimaxime[Jensen, 2001], [Held, Sabanes Bove, 2014].

Reţelele bayesiene aditive pot fi un instrument util în analiza datelor de rezistenţă antimicrobiană, ele fiind aplicate cu succes în identificarea interacţiunilor dintre rezistenţele la antibiotice şi potenţiali factori de risc [Ludwig et. al., 2013], [Hartnack et. al., 2019], [Kratzer et. al. 2020].

Printre limitările metodei ABN în contextul analizei AMR se pot menţiona următoarele:

- tehnica este costisitoare din punct de vedere computaţional întrucât trebuie analizate un număr foarte mare de configuraţii de reţea candidate;
- potenţial de instabilitate numerică în cazul datelor rare şi în care apare fenomenul de separare;
- datele trebuie să fie complete (fără valori lipsă);
- există tendinţa de supraantrenare a modelului în momentul în care trebuie identificată o reţea complexă de interacţiuni dintr-un volum limitat de date; aceasta poate fi redusă folosind metode de tip Markov Chain Monte Carlo.

2.2. Reguli de asociere

Analiza regulilor de asociere [Agrawal et. al. 1993], [Agrawal, Srikant 1994] este o tehnică de învăţare automată ce vizează explorarea seturilor de date de mari dimensiuni cu scopul identificării unor pattern-uri potenţial interesante constând în submulţimi de obiecte "asociate" (i. e., care apar împreună mai frecvent decât ar fi de aşteptat sub ipoteza de independenţă statistică). Vom aminti în continuare câteva concepte şi principii de bază ale acestei tehnici, o prezentare detaliată fiind făcută anterior în Raportul R 5.1.2 Modele şi tehnici computaţionale selectate pentru analiza datelor corelata cu rezistenţa antimicrobiană.

În analiza de asociere datele sunt reprezentate sub formă de liste de tranzacţii. Vom considera $O = \{o_1, o_2, \dots, o_n\}$ o mulţime de obiecte şi $T = \{t_1, t_2, \dots, t_N\}$ o listă de tranzacţii, unde fiecare tranzacţie este o submulţime a lui O înregistrată în setul de date ca urmare a unei acţiuni. Frecvenţa unei mulţimi $A \subset O$ în acest set de date se defineşte ca fiind numărul tranzacţiilor care o conţin:

$$\sigma(A) = card\{t \in T : A \subseteq t\}.$$

Prin urmare, probabilitatea (estimată) ca obiectele din A să apară simultan într-o tranzacţie este $P(A) = \frac{\sigma(A)}{N}$.

O regulă de asociere este o implicaţie de forma $A \to B$, cu $A, B \subset O$ disjuncte. Mulţimile A şi B se numesc antecedentul, respectiv consecventul regulii.

Există numeroase măsuri cu ajutorul cărora poate fi evaluată relevanţa unei reguli de asociere (gradul său de interes pentru utilizator) - a se vedea de exemplu [Tan et. al., 2004], [Geng, Hamilton, 2006]. Între acestea menţionăm:

- suportul: $supp(A \to B) = \frac{\sigma(A \cup B)}{N} = P(A \cap B)$;
- încrederea (confidenţa): $conf(A \to B) = \frac{\sigma(A \cup B)}{\sigma(A)} = P(B|A)$;
- liftul (factorul de interes): $lift(A \to B) = \frac{P(A \cap B)}{P(A) \cdot P(B)}$.

Suportul este proporţia din totalul tranzacţiilor în care A şi B apar împreună. O regulă cu suport mic apare rar în date, deci este posibil ca ea să fie doar o coincidenţă. Încrederea reprezintă proporţia din totalul tranzacţiilor care îl conţin pe A în care apare şi B, i.e., probabilitatea de a-l observa pe B într-o tranzacţie care îl conţine deja pe A – prin urmare, vor fi de interes regulile de asociere cu încredere cât mai mare. Liftul regulii $A \to B$ indică în ce măsură A şi B apar împreună în tranzacţii mai frecvent decât ar fi de aşteptat dacă ele ar fi independente (este cunoscut că, în ipoteza de independenţă, $P(A \cap B) = P(A) \cdot P(B)$).

Scopul în analiza de asociere este acela de a extrage dintr-o bază de date toate regulile cu suport şi încredere mai mari decât anumite valori specificate $minsup$, respectiv $minconf$. Procesul presupune două etape:

- generarea tuturor mulţimilor de obiecte care au suportul mai mare decât $minsup$;
- generarea tuturor regulilor de încredere mai mare ca $minconf$ ce implică mulţimile găsite la pasul anterior.

Analiza regulilor de asociere a fost propusă ca metodă de studiu al pattern-urilor de rezistenţă antimicrobiană (AMR) în lucrare [Cazer et. al. 2019]. În acest context, obiectele considerate sunt rezistenţele la diferite antibiotice, iar setul de tranzacţii este o mulţime de izolate bacteriene având fiecare înregistrate prezenţa sau absenţa fiecărui tip de rezistenţă, urmărindu-se identificarea grupurilor de antibiotice la care apare frecvent rezistenţă concomitentă. O regulă de asociere în acest caz este de tipul "dacă o tulpină bacteriană este rezistentă la antibioticele *A*, atunci ea va fi de asemenea rezistentă la *B*". Un studiu al utilităţii diverselor măsuri de interes pentru reguli în aplicaţiile ce privesc analiza AMR a fost realizat în lucrarea [Zaharia et. al., 2019].

Este important de menţionat că metoda prezintă o serie de avantaje faţă de tehnicile utilizate anterior în analiza AMR (a se vedea [Cazer et. al. 2019]), printre care lipsa ipotezelor distribuţionale asupra datelor şi capacitatea de a gestiona uşor volume mari de date. Există de asemenea câteva limitări de care trebuie să se ţină seama, cele mai importante fiind legate de extragerea regulilor rare (a se vedea [Lopera Gonzales et al, 2019]).

Trebuie avut în vedere că nu întotdeauna regulile extrase pe baza unui eşantion de date reprezintă asocieri reale la nivelul populaţiei, în unele cazuri putând fi vorba despre asocieri false sau simple coincidenţe (corespondent conceptului de eroare de tip I din statistică). Un control asupra ratei descoperirilor false se poate realiza filtrând setul de reguli prin impunerea unui prag minimal de suport *minsup* suficient de mare, o modalitate de estimare a acestuia fiind propusă în [Zaharia et. al., 2019]. Impunând un astfel de prag, regulile extrase beneficiază de un suport statistic corespunzător. Dezavantajul este că un astfel de demers exclude din start şi asocierile reale dar care vizează elemente implicate într-un număr foarte mic de tranzacţii. Spre exemplu, este posibil ca prevalenţa izolatelor dintr-o specie bacteriană rezistente la un antibiotic *A* să fie una extrem de redusă (e.g. în cazul medicamentelor noi), dar toate acestea să fie de asemenea rezistente la antibioticul *B*. O astfel de asociere poate fi una cu implicaţii practice foarte importante, dar datorită suportului foarte scăzut, ea nu va fi detectată. Un fenomen conex este acela de "diluare a suportului" – prin simpla adăugare la setul de date a unor tranzacţii care nu conţin obiectele dintr-o anumită regulă, suportul său va scădea, devenind în final nedetectabilă.O abordare alternativă, bayesiană, în extragerea regulilor de asociere care evită limitările filtrării bazate pe suport a fost propusă în lucrarea recentă [Lopera Gonzales et al, 2019].

2.3. Utilizarea regulilor de asociere în modelarea reţelelor de interacţiuni cu ajutorul ABN

În multe situaţii practice (şi frecvent, în analiza pattern-urilor de rezistenţă antimicrobiană) se pune problema modelării unei reţele complexe de interacţiuni dintr-un volum limitat de date rare, adeseori imperfecte, şi în care relaţiile relevante pot fi mascate de zgomot. Acest fapt poate pune probleme considerabile oricărui algoritm de învăţare. Pentru limitarea acestora, numeroşi autori susţin includerea de informaţii suplimentare care să ghideze procesul de învăţare. Acestea pot fi cunoştinţe anterioare specifice domeniului sau informaţii extrase din date utilizând tehnici complementare.

Aşa cum a fost discutat în Secţiunea 2.1, determinarea structurii optimale a modelului ABN se poate face fie printr-o analiză exhaustivă a tuturor topologiilor de reţele candidate, fie, atunci când numărul structurilor posibile este prohibitiv de mare, printr-un algoritm de căutare euristică ce optimizează o funcţie scor – existând riscul semnificativ al identificării doar a unui optim local. Acest risc, caracteristic algoritmilor de optimizare în general, poate fi limitat: în lucrarea [Wolpert, Macready 1997], Wolpert şi Macready au arătat că utilizarea de informaţie suplimentară (cunoştinţe specifice domeniului) poate oferi un ajutor algoritmului în explorarea spaţiului stărilor unei probleme care să conducă la identificarea de soluţii aproape optimale.

În contextul reţelelor bayesiene, căutarea structurii optimale de reţea poate fi ghidată utilizând topologii preliminare bazate pe informaţii anterioare asupra existenţei unor anumite sub-structuri (a se vedea [Imoto et. al., 2003], [Castelo, Siebes, 1998], [Djebbari, Quackenbush, 2008], [Werhli, Husmeier 2007]). Acestea influenţează căutarea fără a o limita. Spre exemplu, Djebbari şi Quackenbush [Djebbari, Quackenbush, 2008] au arătat că folosirea de informaţii anterioare diverse îmbunătăţeşte semnificativ capabilităţile reţelelor bayesiene de a descoperi reţele de interacţiuni de gene din date de expresie genică, utilizând în acest sens informaţii din literatura biomedicală combinate cu date de interacţiune proteină-proteină.

Pentru integrarea efectivă a surselor de informaţie anterioară într-un model ABN se poate folosi metoda propusă de Werhli şi Husmeier în [Werhli, Husmeier, 2007]. Aceştia oferă o abordare bayesiană pentru inferenţă asupra structurii şi parametrilor reţelei bazată pe eşantionare din distribuţia posterioară utilizând o metodă Markov Chain Monte Carlo (MCMC) ([Madigan, York, 1995], [Friedman, Koller, 2003]).

Reprezentarea informaţiei anterioare extinde o idee din [Imoto et. al., 2003]. Pentru aceasta, considerăm $G = (G_{ij})_{i.j=\overline{1,N}}$ matricea de adiacenţă a grafului (N este numărul total de noduri) şi $B = (B_{ij})_{i.j=\overline{1,N}}$ matricea de informaţie anterioară, unde $B_{ij} \in [0,1]$ reprezintă gradul de convingere prealabilă că există un arc de la i la j. Energia reţelei se defineşte ca

$$E(G) = \sum_{i,j=1}^{N} |B_{ij} - G_{ij}|$$

şi măsoară discrepanţa dintre structura G şi informaţia anterioară B. În cazul în care există K surse de informaţie, fiecare reprezentată de o matrice $B^{(k)} (k = \overline{1,K})$, vom avea K funcţii de energie

$$E_k(G) = \sum_{i,j=1}^{N} |B_{ij}^{(k)} - G_{ij}|, k = \overline{1,K}.$$

Influenţa fiecărei surse de informaţie este controlată de un hiperparametru $\beta_k \in (0, \infty)$; cu cât valoarea lui β_k este mai mare, influenţa informaţiei anterioare relativ la cea a datelor este mai puternică.

Cu ajutorul acestor funcţii de energie se defineşte o distribuţie anterioară Gibbs, probabilitatea anterioară a unei reţele G daţi $\beta_1, \beta_2, \dots, \beta_K$ fiind

$$P(G|\beta_1, \dots, \beta_K) = \frac{\exp(-\sum_{k=1}^{K} \beta_k E_k(G))}{Z(\beta_1, \dots, \beta_K)}$$

cu $Z(\beta_1, \dots, \beta_K) = \sum_{G \in \mathcal{G}} \exp(-\sum_{k=1}^{K} \beta_k E_k(G))$ constantă de normalizare (sumarea se face pe mulţimea tuturor structurilor de reţea posibile). Aceasta este utilizată ulterior în schema de eşantionare MCMC (pentru detalii, a se vedea [Werhli, Husmeier, 2007], Secţiunea 2.4).

În lucrarea [Mureşan et. al., 2020] a fost analizată ideea de a integra ca sursă de informaţie anterioară în procesul de căutare a structurii modelului ABN pattern-urile de asociere extrase din date cu ajutorul regulilor de asociere. Mai exact, gradul de convingere prealabilă B_{ij} asupra existenţei în model a unui arc de la i la j a fost considerat a fi încrederea regulii $i \to j$. Studiul de caz relevă avantaje notabile ale combinării celor două tehnici în analiza datelor de rezistenţă antimicrobiană.Rezultatele sunt prezentate pe larg în Secţiunea 2.5.

2.4. Pachete software utilizate

Metoda ABN a fost implementată în pachetul "abn" din R [Kratzer et. al., 2019]. Scopul acestui pachet este de a implementa o mulţime de funcţii pentru a atribui un scor, selecta, analiza şi returna o reţea bayesiană aditivă. Cele mai importante funcţii ale acestui pachet sunt cele pentru calculul scorurilor, cele pentru algoritmii de căutare exactă şi greedy ale structurii de reţea. Pachetul „abn" mai conţine şi o serie de funcţii auxiliare pentru a manipula modelele ABN.

Scorurile folosite de acest pachet pentru a stabili care este cea mai potrivită reţea pornind de la datele observate sunt BayesianInformation Criterion (BIC), Akaike Information Criterion (AIC) şi Minimum DescriptionLength (MDL), atunci când se foloseşte abordarea estimării prin verosimilitate maximă (abordarea frecvenţionistă). Toate acestea au o componentă care măsoară cât de bine aproximează reţeaua datele şi o componentă care penalizează pentru complexitate. Un al patrulea scor este verosimilitatea marginală, specifică abordării bayesiene.

Funcţiile din "abn" oferă posibilitatea de a elimina sau include din start anumite arce, astfel se pot introduce cunoştinţe de specialitate în model şi se poate configura pentru fiecare nod o mulţime de variabile părinte valide.

Pachetul "abn" conţine trei tipuri de funcţii:

- principale, folosite pentru a determina modelul ABN (atribuirea unui scor, determinarea stucturii şi estimarea parametrilor);
- auxiliare pentru analiză (care reprezintă grafic reţeaua determinată, estimează puterea arcelor şi compară două structuri diferite);
- auxiliare pentru simulare (simulează DAG-uri şi date ABN).

O analiză ABN tipică implică folosirea succesivă a trei funcţii principale din pachet: *buildscorecache()*, *mostprobable()* şi respectiv *fitabn()*. În acest fel utilizatorul are posibilitatea de a modifica diverşi parametri ai funcţiilor pentru a rezolva diverse probleme întâlnite în etapa de învăţare a modelului. Aceste funcţii au atât o implementare bayesiană, cât şi una frecvenţionistă (folosind estimarea prin verosimilitate maximă), care pot genera rezultate diferite. În apelarea lor este necesară specificarea setului de date,

distribuţia variabilelor (nodurilor) şi o limită superioară pentru complexitatea reţelei (un număr maxim de părinţi pentru noduri).

Funcţia *buildscorecache()* determină mai întâi o listă validă în care apar toate combinaţiile posibile de părinţi pentru fiecare nod, ţinând seama de opţiunea utilizatorului de a include sau elimina anumite arce. Aceasta iterează apoi lista de scoruri ataşată fiecărui nod pentru a calcula scorul fiecărei reţele candidate. În abordarea bayesiană, funcţiile *buildscorecache()* şi *fitabn()* (cea din urmă este un wrapper pentru prima) determină o regresie bayesiană folosind următoarele distribuţii apriori: distribuţii gaussiene slab informative cu media 0 şi varianţa 1000, pentru fiecare parametru al regresiei, şi distribuţii Gamma difuze cu ambii parametri egali cu 0.001 pentru parametri de precizie ai nodurilor gausiene folosind metoda de aproximare integrată Laplace (INLA) [Gómez-Rubio, 2020]. În abordarea frecvenţionistă, funcţia *buildscorecache()* foloseşte un algoritm Iterated Reweigthed Least Squares [Faraway, 2016], în funcţie de distribuţia nodurilor. Dacă variabilele au distribuţie binomială, se determină un model de regresie logistică.

Pentru determinarea structurii reţelei au fost implementaţi doi algoritmi: căutarea exactă şi euristică. Primul dintre acestea este implementat de funcţia *mostprobable(),* iar cel de-al doilea de funcţiile *searchHeuristic()* şi *searchHillclimber().* Căutarea exactă a structurii modelului ABN foloseşte abordarea introdusă de Koivisto şi Sood [Koivisto, Sood, 2004] (căutarea pe ordine şi nu pe spaţiul tuturor DAG-urilor posibile). Funcţia *mostprobable()* necesită o listă de scoruri calculate anterior pentru fiecare nod (dată de *buildscorecache()* de exemplu), scorul folosit şi o distribuţie anterioară a structurii. Căutarea exactă nu poate fi făcută pentru mai multe de 20 de variabile, dar are performanţe mai bune decât căutarea folosind metoda Markov Chain Monte Carlo sau alte abordări probabiliste, aceasta determinând DAG-ul optimal global.

Funcţia *searchHeuristic()* implementează algoritmi de căutare euristici cum ar fi "hill-climber", "Tabu" şi "simulated annealing". Aceasta necesită de asemenea o listă a scorurilor calculate anterior pentru fiecare nod, scorul folosit şi câteva argumente ce depind de metoda folosită. DAG-ul returnat este doar local optimal.

Funcţia *fitabn()* atribuie un scor unei reţele date şi necesită un DAG valid, setul de date şi lista distribuţiilor nodurilor. Aceasta returnează lista scorurilor pentru fiecare nod, parametrii estimaţi ai modelului ABN, abaterile standard ale acestora şi p-valorile asociate.

Alte funcţii auxiliare precum *plotabn()* şi *tographviz()* sunt utile în reprezentarea grafică a modelelor ABN. Prima dintre acestea permite vizualizarea DAG-ului având valorile parametrilor modelului pe fiecare arc, grosimea acestora putând fi proporţională cu intensitatea legăturii dintre variabile.

Odată ce s-a determinat un DAG optimal local sau global, se poate stabili dacă modelul găsit este supraantrenat folosind metoda boostrapping: se generează DAG-uri cu un grad mare de suport (scor mare) cu ajutorul unui DAG dat şi se determină ce parte a structurii selectate este susţinută de către date. Acest lucru se poate realiza cu ajutorul funcţiei *mcmcabn()* din pachetul R cu acelaşi nume [Kratzer, Furrer, 2019].

Funcţiile din "mcmcabn" rezolvă următoarele probleme:

- selectează structura cea mai probabilă folosind o listă de scoruri calculate anterior;
- controlează supraantrenarea reţelei;

- eşantionează mulţimea structurilor cu cele mai mari scoruri.

Pachetul "mcmcabn" necesită specificarea unei liste de scoruri calculate anterior (determinate de exemplu folosind funcţia *buildscorecache()*), apoi funcţia *mcmcabn()* generează eşantioane (DAG-uri) folosind distribuţia posterioară a DAG-urilor cu scorurile cele mai mari, care sunt reprezentative pentru date.

Pentru a realiza acest lucru, utilizatorul trebuie să specifice o distribuţie iniţială a DAG-urilor şi a metodă de parcurgere a spaţiului de grafuri posibile. Distribuţiile iniţiale sunt: Koivisto (o distribuţie neinformativă), Ellis sau o distribuţie definită de utilizator; iar algoritmii de parcurgere a spaţiului DAG-urilor sunt: Monte Carlo Markov Chain Model Choice ($(MC)^3$) unde metoda aleasă este Metropolis-Hasting, "new edge reversal" propusă de Grzegorczyk şi Husmeier [Grzegorczyk, Husmeier, 2008] şi "Markov blanket resampling" propusă de Su şi Borsuk [Su, Borsuk, 2016].

Cea mai importantă funcţie din acest pachet este *mcmcabn()* care generează eşantioane de date folosind distribuţia posterioară a DAG-urilor. Cu aceste noi date generate se pot determina structuri optimale local sau global sau se pot calcula probabilităţi de existenţă sau absenţă a unor arce sau a unei părţi a structurii cu funcţia *query().* O altă funcţie din acest pachet este *summary(),* cu ajutorul căreia se pot returna diagnostice pentru metoda Markov Chain Monte Carlo şi câteva metrici descriptive.

În cazul în care numărul datelor nu este mare, cum se întâmplă adesea în epidemiologie, distribuţia apriori considerată joacă un rol important în determinarea modelului ABN. Astfel este important ca aceasta să fie bine aleasă. Funcţiile din pachetul "mcmcabn" lasă utilizatorului opţiunea de a stabili o distribuţie apriori.

Trei astfel de distribuţii sunt implementate în pachetul "mcmcabn". Parametrul "prior.choice" al funcţiei *mcmcabn()* determină tipul de distribuţie ales pentru fiecare nod, dată fiind o mulţime de părinţi ai acestuia. Un tip de distribuţie ce poate fi folosit este Koivisto introdus în [Koivisto, Sood, 2004], care presupune că probabilitatea anterioară a mulţimilor de părinţi cu acelaşi număr de elemente este identică. Mai exact, distribuţia Koivisto este dată de:

$$P(G) = \frac{1}{z}\prod_{n=1}^{N}\binom{N-1}{|G_n|}^{-1}$$

unde N este numărul de noduri, $|G_n|$ este numărul de părinţi ai nodului n, iar z este o constantă normalizatoare. Această distribuţie favorizează mulţimi de părinţi cu un număr foarte mic sau foarte mare de elemente, ceea ce constituie un dezavantaj. Pentru a o folosi, trebuie ca parametrul "prior.choice" să aibă valoarea 2.

Atunci când "prior.choice" are valoarea 1, se foloseşte o distribuţie neinformativă în care toate combinaţiile de părinţi au aceeaşi probabilitate.

Dacă valoarea parametrului "prior.choice" este 3, se foloseşte o distribuţie definită de utilizator, specificată de parametrul "prior.dag" al funcţiei *mcmcabn().* Acesta este o matrice pătratică şi având ordinul egal cu numărul de noduri, în care valorile definesc încrederea utilizatorului că există arcul respectiv în modelul ABN. Un hiperparametru ce defineşte încrederea globală în distribuţia anterioară specificată este dat de parametrul

"prior.lambda". Matricea se defineşte în stilul abordării lui Werhli şi Husmeier [Werhli, Husmeier, 2007] (a se vedea secţiunea 2.3).

2.5. Studiu de caz

Această secţiune se bazează pe rezultatele din lucrarea [Mureşan et. al, 2020].

2.5.1. Prezentarea setului de date

Baza de date National AntimicrobialResistance Monitoring System for Enteric Bacteria (NARMS)a fost înfiinţată în 1996 în Statele Unite ale Americii şi este un sistem naţional de monitorizare a rezistenţei antimicrobiene a unor bacterii precum Salmonella enterica, Campylobacter şi Escherichia coli regăsite la subiecţi umani, animale domestice, carne şi produse din carne [NARMS, 2019].

Setul de date folosit în această analiză a fost extras din NARMS şi constă din 329 izolate de Escherichia Coli provenind de la subiecţi umani şi colectate între anii 1996 şi 2016. Fiecare izolat a fost clasificat ca rezistent sau susceptibil, pe baza pragurilor de concentraţie minimă inhibitoare, la următoarele antibiotice: *ampicilină* (AMP), *amoxicilină – acid clavulanic* (AUG), *ceftriaxonă* (AXO),*cloramfenicol* (CHL), *ciprofloxacină* (CIP), *acid nalidixic*(NAL), *gentamicină* (GEN), *streptomicină* (STR),*trimetoprim – sulfametoxazol* (COT) şi*tetraciclină*(TET). Astfel, fiecare antibiotic este reprezentat de o variabilă Bernoulli, având valorile 0 (susceptibil) sau 1 (rezistent).

Tabelul de mai jos arată prevalenţa rezistenţei izolatelor testate la fiecare antibiotic.

	AMP	AUG	AXO	CHL	CIP	COT	GEN	NAL	STR	TET
Nr. de izolate	106	14	13	60	10	48	15	60	147	200
%	32.22	4.26	3.95	18.24	3.04	14.59	4.56	18.24	44.68	60.79

Tabelul 1. Prevalenţa rezistenţei la cele 10 antibiotice studiate

2.5.1. Metodologia de analiză

Analiza statistică a datelor s-a făcut folosind pachetele R "abn" şi "mcmcabn" descrise anterior. A fost mai întâi determinată structura de reţea folosind o căutare exactă, scorul utilizat fiind BIC. Intrucât setul de date prezintă probleme semnificative de separare şi sparsitate, s-a utilizat metoda Markov Chain Monte Carlo (MCMC) pentru limitarea supraantrenării modelului. Au fost generate patru lanţuri pornind de la DAG-ul global optimal determinat anterior, fiecare având o lungime de 50000 de paşi, o etapă de burn-in de 5000 de paşi şi un pas de filtrare de 100 pentru evitarea autocorelaţiei. Convergenţa lanţurilor a fost evaluată cu ajutorul graficelor diagnostic şi a statisticilor de convergenţă Gelman-Rubin [Gelman, Rubin, 1992]. Media acestor patru lanţuri a condus la constituirea DAG-ului de consens majoritar, eliminându-se arcele cu suport mai mic de 50%.

În următoarea etapă au fost încorporate în modelul ABN cunoştinţele suplimentare extrase cu ajutorul regulilor de asociere. Mai exact a fost considerată matricea de informaţie anterioară B având elementele B_{ij} egale cu încrederea regulii i → j(au fost utilizate toate

regulile având exact un obiect în antecedent şi unul în consecvent). Hiperparametrul β a fost fixat la valoarea 1. Din nou s-a procedat la generarea de lanţuri şi determinarea DAG-ului de consens majoritar folosind MCMC, ca mai sus.

2.5.2. Rezultate

Modelul ABN optimal global al asocierilor de rezistenţă conţine 16 arce. Acesta a fost folosit ca DAG iniţial pentru metoda Markov Chain Monte Carlo pentru a limita problema supraantrenării conducând la DAG-ul de consens majoritar prezentat în figura de mai jos.

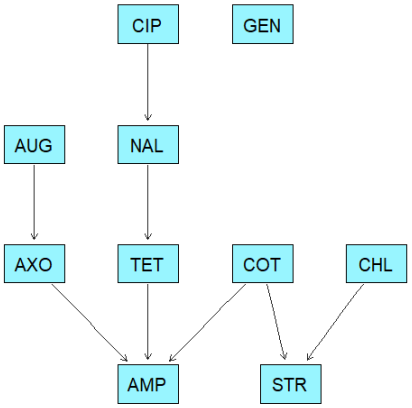


Figura 1. Modelul ABN după limitarea supraantrenării folosind MCMC, având DAG-ul optimal global ca punct de start

Doar 8 din arcele modelului optimal global au fost regăsite în DAG-ul din Figura 1, indicând existenţa supraantrenării reţelei iniţiale. Tabelul de mai jos prezintă valorile raportului de şanse (OR) pentru fiecare arc de tipul ”părinte→copil”, intervalele de încredere de tip Wald şi suportul asociat (procentajul grafurilor generate de metoda MCMC ce conţin acest arc). Mai sunt prezentate valorile măsurilor suport, încredere şi lift pentru regulile de asociere corespunzătoare fiecărui arc.

Tabelul 2. Coeficientul OR, intervalele de încredere 95% pentru acesta, suportul arcului, suportul, încrederea şi liftul regulilor asociate pentru modelul din Figura 1

Arc (p→c)	**OR (95% CI)**	**Suportul arcului (%)**	**Suportul regulii**	**Încrederea regulii**	**Lift-ul regulii**
COT→AMP	9.03 (4.37, 18.64)	74.75	0.09	0.65	2.00
AXO→AMP	561.16 (0.48, 6.51e+05)	74.25	0.04	1.00	3.10
TET→AMP	0.29 (0.17, 0.52)	55.24	0.15	0.25	0.78
CIP→NAL	47.47 (5.83, 386.53)	54.93	0.03	0.90	4.94
NAL→TET	0.23	54.55	0.06	0.32	0.52

	(0.12, 0.41)				
COT→STR	4.66 (2.17, 10.02)	52.87	0.12	0.79	1.77
CHL→STR	4.44 (2.28, 8.64)	52.56	0.14	0.77	1.72
AUG→AXO	259.82 (51.07, 1321.87)	51.12	0.03	0.71	18.08

Asocieri pozitive puternice au fost găsite între rezistenţele la COT şi AMP, CIP şi NAL, COT şi STR, CHL şi STR, şi AUG şi AXO. De asemenea, rezistenţele la antibioticele AXO şi AMP apar a fi puternic asociate pozitiv, având o valoare a coeficientului OR de 561. Totuşi intervalul de încredere de 95% corespunzător este foarte mare şi conţine valoarea 1, ceea ce ar însemna că această asociere nu este semnificativă statistic. Valorile măsurilor regulii AXO→AMP ajută la deducerea unei posibile explicaţii pentru acest lucru: încrederea este 1, ceea ce înseamnă că toate izolatele rezistente la AXO din setul de date sunt rezistente şi la AMP, iar liftul este 3.1 – aşadar asocierea este într-adevăr relevantă, iar intervalul de încredere mare este o consecinţă a instabilităţii numerice datorate separării datelor.

În lucrarea [Zaharia et. al. 2019] a fost folosit un prag de suport minim de 0.06 pentru a limita rata descoperirilor false la 5%. Aşa cum s-a menţionat anterior, în acest mod se pot elimina regulile rare împreună cu asocierile întâmplătoare. Astfel, asocierile dintre rezistenţele la AXO şi AMP, CIP şi NAL, AUG şi AXO sunt foarte puternice, dar apar rar deoarece rezistenţele individuale la antibioticele AXO, AUG şi CIP sunt rare (a se vedea Tabelul 1). Aşadar aceste asocieri au fost eliminate atunci când s-a folosit metoda determinării regulilor de asociere, dar au fost descoperite de ABN.

În plus, s-au descoperit asocieri negative între rezistenţele la TET şi AMP, şi NAL şi TET, ale căror reguli asociate au încrederea şi liftul mai mici ca 1.

Atunci când s-a folosit abordarea descrisă în Secţiunea 2.3 de includere a cunoştinţelor extrase din regulile de asociere ca distribuţie anterioară a arcelor, a fost obţinut modelul ABN prezentat în Figura 2 şi Tabelul 3.

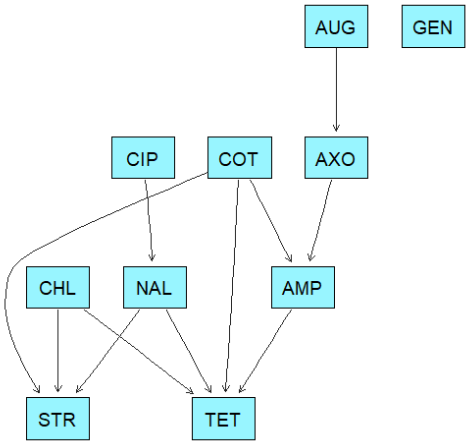


Figura 2. Modelul ABN obţinut folosind metoda MCMC, având ca punct de start un DAG vid şi distribuţia anterioară pentru fiecare arc dată de încrederea regulii asociate

Tabelul 3. Coeficientul OR, intervalele de încredere 95% pentru acesta, suportul arcului, suportul, încrederea şi liftul regulilor asociate pentru modelul din Figura 2.

Arc (p→c)	OR (95%CI)	Suportul arcului (%)	Suportul regulii	Încrederea regulii	Lift-ul regulii
COT→TET	8.17 (3.19, 20.92)	79.70	0.12	0.81	1.33
AMP→TET	0.18 (0.10, 0.32)	79.17	0.15	0.47	0.78
NAL→TET	0.13 (0.07, 0.26)	77.82	0.06	0.32	0.52
AXO→AMP	720.54 (0.22, 2.41e+06)	74.35	0.04	1.00	3.10
COT→STR	5.47 (2.50, 11.99)	70.72	0.12	0.79	1.77
CHL→STR	4.1 (2.10, 7.98)	65.67	0.14	0.77	1.72
COT→AMP	5.93 (3.11, 11.32)	65.43	0.09	0.65	2.00
CIP→NAL	47.47 (5.83, 386.53)	61.27	0.03	0.90	4.94
NAL→STR	0.31 (0.15, 0.64)	58.27	0.05	0.27	0.60
AUG→AXO	259.82 (51.07, 1321.87)	52.00	0.03	0.71	18.08
CHL→TET	3.46 (1.55, 7.72)	51.00	0.15	0.82	1.34

Pe lângă asocierile descoperite anterior, această abordare a permis descoperirea unor noi legături, şi anume asocieri pozitive între rezistenţele la COT şi TET, CHL şi TET, precum şi o legătură semnificativă negativă între rezistenţele la NAL şi STR. Primele două asocieri au fost regăsite şi în [Zaharia, 2019], folosind doar analiza regulilor de asociere, şi reprezintă pattern-uri documentate clinic [Batard et. al. 2016, Tadesse et. al. 2012].

2.5.3. Discuţie

Analiza de mai sus arată avantajele combinării metodei extragerii regulilor de asociere şi reţelelor bayesiene aditive pentru descoperirea pattern-urilor de rezistenţă antimicrobiană. S-a putut observa că regulile de asociere pot fi folosite atât cu rol de cunoştinţe anterioare pentru a determina modelul ABN, cât şi ca suport în interpretarea rezultatelor date de reţeaua obţinută. Astfel, folosind împreună cele două metode se pot

obţine informaţii atât despre intensitatea, cât şi despre frecvenţa asocierilor, depăşindu-se limitările specifice fiecărei metode.

Aşa cum s-a putut observa anterior, utilizând încrederea regulilor ca distribuţie anterioară a acestor asocieri în modelul ABN, au fost descoperite pattern-uri suplimentare cu relevanţă clinică.

3. Clasificare multi-etichetă în analiza rezistenţei la antibiotice

3.1. Specificul clasificarii multi-etichetă

Predicţia rezistenţei simultane la mai multe antibiotice poate fi interpretată ca o problemă de clasificare multi-etichetă (Multi-Label Classification – MLC). Spre deosebire de clasificarea tradiţională, în care fiecare dată de intrare aparţine uneia dintre clase, în clasificarea multi-etichetă fiecare dintre date poate să aparţină simultan mai multor clase. Aceasta face ca problema de clasificare să devină mai dificilă, iar modelele tradiţionale de clasificare (arbori de decizie, clasificatori bazaţi pe instanţe sau pe cel mai apropiat vecin, clasificatori probabilişti de tip Bayesian sau clasificatori bazaţi pe vectori suport) nu pot fi aplicaţi direct. In acest context, în ultimii ani au fost propuse o serie de metode de clasificare specifice pentru MLC (Herrera et al., 2016). In contextul procesării datelor biologice, clasificarea multi-etichetă este cel mai frecvent utilizată în predicţia funcţiilor biologice ale genelor, întrucât fiecare genă corespunde adesea mai multor funcţii.

3.2. Modele de clasificare multi-etichetă

Tehnicile pentru construirea modelelor de clasificare multi-etichetă pot fi grupate în trei categorii principale:

- *Tehnici bazate pe transformarea problemei*. Aceste tehnici presupun transformarea problemei de clasificare multi-etichetă în probleme de clasificare binară sau multi-clasă. Cele mai populare abordări sunt(Herrera et al., 2016):
  - *Binary Relevance - BR*: în cazul a k etichete, problema este transformată în k probleme de clasificare binară; dezavantajul major al acestei abordări este că nu ţine cont de eventualele corelaţii între etichete.
  - *Label Powerset - LP*: fiecare dintre vectorii binari corespunzători etichetelor este interpretat ca fiind reprezentarea în baza doi a unei etichete, astfel că problema iniţială este transformată într-una de clasificare multi-clasă cu $m \leq 2^k$; problema principală este că în construirea modelului sunt luate în considerare doar configuraţii care sunt prezente în setul de date.
  - *Classifier Chain– CC(Read, 2009)*: problema iniţială este redusă la k probleme de clasificare binară, ca la metodele de tip BR, cu diferenţa că se construiesc clasificatori corelaţi (rezultatul de la clasificatorul curent este adăugat ca variabilă de intrare pentru clasificatorul următor). Principalul dezavantaj al acestei abordări este faptul că performanţa clasificării este influenţată de ordinea de înlănţuire a clasificatoarelor.

Fiecare dintre aceste strategii poate fi combinată cu oricare dintre clasificatorii binari sau multi-clasă tradiţionali fără a necesita modificarea acestora din urmă.

- *Tehnici bazate pe adaptarea metodelor*. Presupun modificarea metodelor de clasificare binară sau multi-clasă pentru a încorpora cazul multi-etichetă. Câteva dintre familiile de clasificatori pentru care au fost dezvoltate variante multi-etichetă sunt:
  - *Arbori de decizie*. Adaptarea presupune modificarea modului de calcul al criteriilor (entropie, câştig informaţional etc) utilizate în selecţie atributului de ramificare şi în asignarea etichetelor finale.
  - *Metode bazate pe instanţe (k Nearest Neighbour)*.Singura componentă din kNN care trebuie modificată este strategia de votare în stabilirea claselor din care face parte o dată de intrare. In ML-kNN (Zhang, 2007)strategia de votare se bazează pe estimarea, pentru fiecare etichetă, a probabilităţii ca data de intrare să aparţină clasei respective. Estimarea utilizează regula lui Bayes folosind frecvenţe relative estimate pa baza setului de vecini.
  - *Reţele neuronale*. Extinderea reţelelor neuronale de tip perceptron multiplu (MLP) sau dintre cele bazate pe funcţii cu simetrie radială (RBF) pentru clasificarea multi-etichetă este naturală întrucât la nivelul de ieşire se colectează un vector de valori. Principala dificultate în acest caz este stabilirea pragurilor aferente diferitelor clase. In contextul arhitecturilor cu structură adâncă au fost propuse recent soluţii privind determinarea adaptivă a pragurilor (Ghoshal et. al, 2020).
  - *Clasificatori bazaţi pe vectori suport (Support Vector Machines – SVM).* O variantă de extindere a ideii de clasificare bazată pe funcţii suport este cea din (Elisseeff, & Weston, 2001)se bazează pe definirea unei funcţii de eroare bazată pe ierarhizarea răspunsurilor returnate de m clasificatoare, m fiind numărul de configuraţii multi-eticheta posibile.
- *Tehnici de tip ansamblu.* Se bazează pe ideea de a agrega diferite modele folosind strategii de bagging, boosting sau stacking. Exemple de tehnici de tip ansamblu pentru clasificarea multi-etichetă sunt:
  - *Ensemble of Classifier Chains - ECC.* Se bazează pe utilizarea mai multor lanţuri de clasificatoare caracterizate prin altă ordine de înlănţuire a clasificatoarelor. Prezintă avantajul că e eliminată sensibilitatea la ordinea de înlănţuire a clasificatoarelor (cum se întâmplă la CC).
  - *Ensemble of Pruned Sets* – EPS. Se bazează pe construirea mai multor clasificatoare independente bazate pe strategie de tip Label Powerset folosind un subset aleator de antrenare. Se caracterizează prin faptul că utilizează o schemă de votare care permite predicţia unor combinaţii de etichete care nu sunt prezente în setul iniţial de date.
  - *Random k-Labelsets – RAkEL*. Se bazează pe generarea unor m subseturi aleatoare de câte k etichete şi antrenarea unui clasificator multi-clasă pentru fiecare subset.

3.3. Estimarea incertitudinii în modele de clasificare

Modelele de clasificare de tip black-box, cum sunt reţelele neuronale, deşi sunt performante din perspectiva calităţii clasificării au un dezavantaj important: rezultatele nu sunt uşor de interpretat iar, spre deosebire de modelele bazate pe regresie, nu permit estimarea directă a gradului de incertitudine(Lakshminarayanan et al., 2017). Estimarea gradului de incertitudine este importantă în particular în cazul modelelor aplicate în domenii în care costul unei erori în răspunsul furnizat de model este mare cum este domeniul controlului autonom al vehiculelor şi cel medical(Amodei et al., 2016). În ultimii ani s-au înregistrat progrese notabile în ceea ce priveşte estimarea incertitudinii predicţiei, majoritatea bazate pe o abordare bayesiană. Antrenarea modelelor de clasificare folosind tehnici tradiţionale de statistică Bayesiană necesită un efort computaţional semnificativ, astfel că în contextul antrenării reţelelor neuronale cu structură adâncă a devenit populară o strategie bazată pe tehnica de dezactivare aleatoare a unor parametri cunoscută sub denumirea Monte Carlo Dropout (Gal & Gahramani, 2016). Tehnica a fost aplicată cu succes în probleme de clasificare binară şi multi-clasă, însă doar foarte recent a fost testată şi în contextul multi-etichetă (Ghoshal et. al, 2020), ideea fundamentală fiind dezactivarea unor conexiuni din cadrul reţelei în loc de dezactivarea unor unităţi funcţionale în întregime. In contextul rezistenţei la antibiotice sau antivirale au fost raportate câteva rezultate care ilustrează faptul tehnicile de clasificare multi-etichetă permit obţinerea unei performanţe mai bune în contextul predicţiei rezistenţei multiple folosind informaţii privind rezistenţa încrucişată în cazul HIV-1 (Heider, 2013)(Riemenschneider, 2016).

3.4. Studiu de caz

Scopul acestui studiu de caz a fost analiza comparativă a mai multor tehnici de clasificare multi-etichetă în contextul predicţiei rezistenţei multiple la antibiotice a unei tulpini de E. Coli pornind de la prezenţa unor mutaţii într-o serie de gene implicate în dezvoltarea rezistenţei la antimicrobiene. In acest context a fost utilizat un set de date public, accesibil în bazele de date de la NCBI.

Setul de date constă din 363 de înregistrări, fiecare conţinând informaţii genotipice referitoare la 140 de gene(eventuale mutaţii prezente) şi informaţii fenotipice despre rezistenţa/susceptibilitatea la 34 de antibiotice. Tipurile de mutaţii incluse în setul de date sunt:(1) mutaţie punctuală (point-type mutation); (2) mutaţie parţială (partial mutation); (3) mutaţie parţială la final de contig (partial end of contig); (4) eroare în procesul de translaţie(mistranslation) precum şi cazul în care (5) nu e prezentă nicio mutaţie (complete gene - no mutation).Pe de altă parte pentru fiecare dintre cele 34 de antibiotice este specificat dacă E.Coli este rezistentă/susceptibilă la antibioticul respectiv sau dacă nu există informaţie în acest sens. In etapa de pre-procesare a setului de date au fost excluse liniile respectiv coloanele care nu conţineau informaţii (nicio informaţie genotipică sau fenotipică). Ca rezultat al acestei filtrări au rămas în setul de date 335 de înregistrări şi 30 de antibiotice.

Modele şi metode validate: reţele Bayesiene, reguli de asociere şi metode de clasificare multi-etichetă

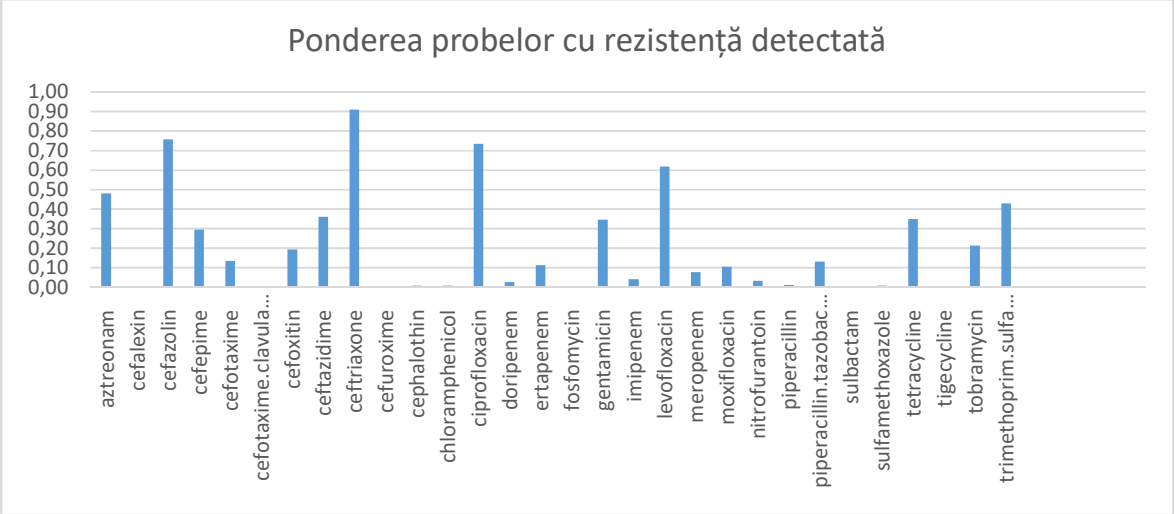


Figura 3. Ponderea probelor pentru care a fost detectată rezistenţă la fiecare dintre cele 30 de antibiotice (setul de date EColi – NCBI AMR)

In scopul construirii unor modele de predicţie multi-etichetă au fost parcurse mai multe etape:

Etapa 1. Analiza caracteristicilor datelor din perspectiva gradului de multiplicitate a etichetelor (multilabelness). Metricile utilizate în acest scop au fost:

- Cardinalitate – numărul mediu de etichete pe instanţă (înregistrare)
- Densitate –cardinalitatea împărţită la numărul de clase
- Procentul înregistrărilor cu o singură etichetă (probe pentru care s-a înregistrat rezistenţă la un singur antibiotic.

Valorile obţinute pentru setul analizat sunt: cardinalitate = 7.80, densitate = 0.229, procent etichetă unică = 1.49%, ceea ce confirmă faptul că problema de clasificare necesită utilizarea unor tehnici specifice pentru MLC.

Etapa 2. Compararea mai multor tehnici de clasificare multi-etichetă. Tehnicile selectate sunt:

- BR (Binary Relevance) + RF (Random Forest) - tehnica transformării datelor combinată cu un clasificator de tipul „ansamblu aleator de arbori de decizie”
- BR (Binary Relevance) + kNN (k Nearest Neighbour) – tehnica transformării datelor combinată cu un clasificator de tipul “cel mai apropiat vecin”
- CC (Classifier Chain) + RF – clasificatori de tip RF înlănţuiţi (Read, 2009)
- ML – kNN (Multilabel k Nearest Neighbour) – varianta extinsă pentru clasificare multi-etichetă a clasificatorilor de tipul “cel mai apropiat vecin” (Zhang, 2007)

Pentru compararea rezultatelor au fost utilizate două măsuri de performanţă:

- Hamming Loss – bazat pe calculului distanţei Hamming dintre vectorul binar corespunzător clasificării corecte şi cel produs de către clasificator;

- F-measure - care se bazează pe agregarea a două altor măsuri (recall şi precision) calculate în raport

Rezultatele comparative corespunzătoare acestor măsuri sunt ilustrate în Figurile 3 şi 4.

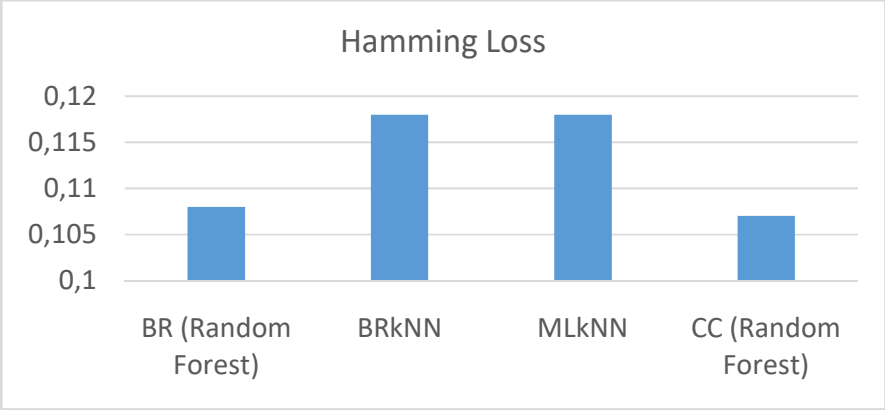


Figura 4. Comparaţie între performanţele algoritmilor de clasificare multi-etichetă. Criteriu: Hamming Loss (valori mai mici indică o performanţă mai bună)

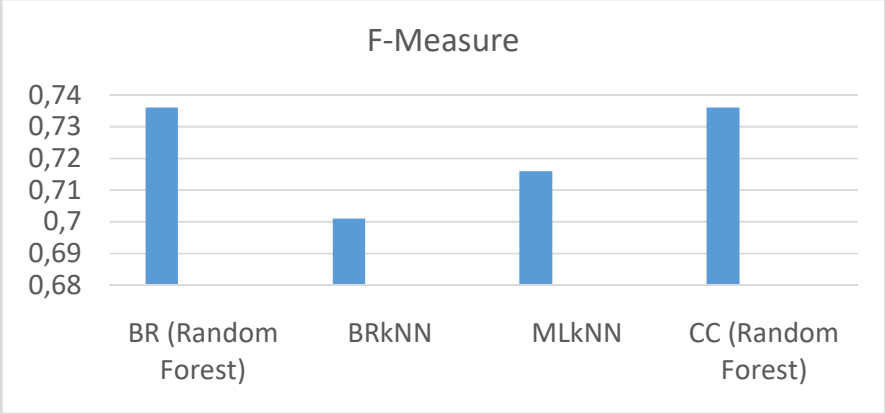


Figura 5. Comparaţie între performanţele algoritmilor de clasificare multi-etichetă. Criteriu: F-measure (valori mai mari indică o performanţă mai bună)

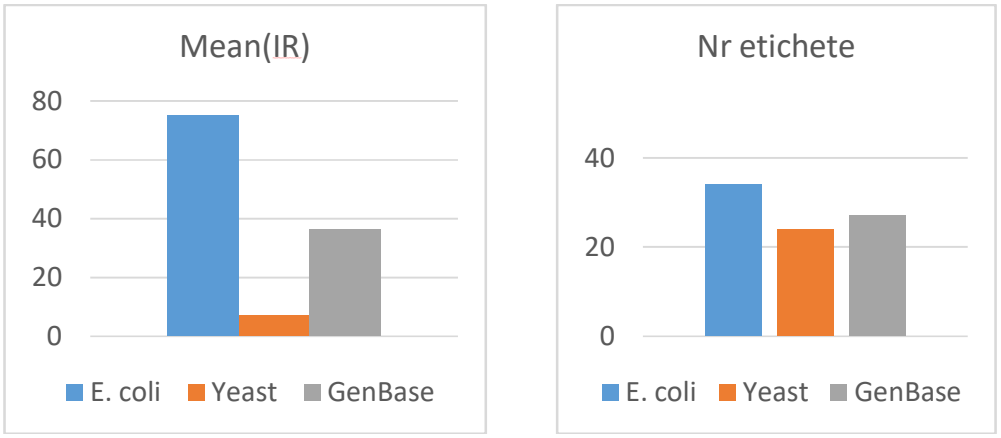
Valorile obţinute şi ilustrate în Fig. 4 şi 5 sunt:

- 0.107 pentru BR combinat cu Random Forest,
- 0.117 pentru BR combinat cu kNN,
- 0.117 pentru MLkNN
- 0.106 pentru CC combinat cu Random Forest.

Aceste rezultate sugerează că cea mai bună performanţă este obţinută prin utilizarea unui clasificator de tip RF şi faptul că modalitatea de extindere a metodelor pentru cazul multi-etichetă nu influenţează în mod semnificativ performanţa.

Etapa 3. Abordarea problemei debalansării. Absenţa unui echilibru între numărul de exemple corespunzător diferitelor clase creează dificultăţi în construire unor modele eficiente. Spre deosebire de cazul clasificării uni-etichetă pentru care este uşor de analizat gradul de debalansare a unui set de date, în cazul multi-etichetă este necesară utilizarea unor indicatori specifici (Charte, 2015):

- *Imbalance Ratio - IRLbl.* IRLbl se calculează independent pentru fiecare etichetă, ia valoarea 1 pentru cea mai frecventă etichetă şi valori din ce în ce mai mari pentru etichete mai puţin frecvente.
- *Average Imbalance Ratio - MeanIR.* Este media valorilor IRLbl corespunzătoare tuturor etichetelor. Cu cât această valoare este mai mare cu atât este mai debalansat setul de date. Pentru setul de date analizat valoarea corespunzătoare este semnificativ mai mare decât a altor seturi de date (Figura 6a).
- *Coefficient of Variation for the average imbalance ratio – CVIR.*



(a) (b)

Figura 6. Valoarea indicatorului de debalansare (a) şi a numărului total de etichete pentru trei seturi de date

Aplicând o tehnică de echilibrare a setului de date de tipul SMOTE (Synthetic Minority Oversampling Technique) se poate obţine o uşoară îmbunătăţire a performanţei.

4. Concluzii

Reţelele Bayesiene aditive sunt instrumente utile în analiza interacţiunilor dintre entităţile care sunt implicate în mecanismele rezistenţei antimicrobiene şi în identificarea unor tipare de interacţiune. Rezultatele obţinute în studiul de caz descris în sect. 2.5. ilustrează faptul că informaţiile încorporate în regulile de asociere extrase din date pot fi utilizate atât sub forma de cunoştinţe apriori pentru ghidarea procesului de construire a reţelei de interacţiuni cât şi în interpretarea finală a reţelei construite.

Pe de altă parte, studiul de caz preliminar descris în sect. 3.4. ilustrează faptul că metodele de clasificare multi-etichetă pot fi utilizate în predicţia rezistenţei multiple şi în construirea de modele care descriu influenţa diferitelor categorii de mutaţii ce intervin în gene responsabile în dezvoltarea rezistenţei la antibiotice.

Bibliografie:

1. (Agrawal et al., 1993) R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data, Washington, DC, May 1993, pp. 207-216.
2. (Agrawal & Srikant, 1994) R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, 1994, pp. 487-499.
3. (Amodei et al., 2016) Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety," http://arxiv.org/abs/1606.06565, 2016, arXiv: 1606.06565.
4. (Batard et al., 2016) E. Batard, M. Lefebvre, G. Ghislain Aubin, N. Caroff, S. Corvec, "High prevalence of cross-resistance to fluoroquinolone and cotrimoxazole in tetracyclineresistant Escherichia coli human clinical isolates", Journal of Chemotherapy, 2016, DOI: 10.1179/1973947815Y.0000000038.
5. (Castelo & Siebes, 1998) R. Castelo, A. P. J. M. Siebes, Priors on network structures: Biasing the search for Bayesian networks, Centrum voor Wiskunde informaticae, 1998.
6. (Cazer et al., 2019) C. L. Cazer, M. A. Al-Mamun, K. Kaniyamattam, W. J. Love, J. G. Booth, C. Lanzas, Y. T. Gröhn, Shared multidrug resistance patterns in chicken-associated Escherichia coli identified by association rule mining, Frontiers in Microbiology, 10, 2019, 687.
7. (Charte, 2015) F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera. Addressing imbalance in multilabel classification: measures and random resampling algorithms. In: Neurocomputing, 2015, pp. 163, 3–16
8. (Djebbari& Quackenbush, 2008) A. Djebbari, J. Quackenbush. "Seeded Bayesian Networks: constructing genetic networks from microarray data", BMC systems biology, 2, 2008, pp. 1-13.
9. (Elisseeff,& Weston, 2001) Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances inNeural Information Processing Systems, vol. 14, pp. 681–687. MIT Press (2001)
10. (Faraway, 2005) J. J. Faraway, Linear models with R, Chapman & Hall, CRC Texts in Statistical Science Series, 2005.
11. (Friedman & Koller, 2003) N. Friedman, D. Koller, "Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks", Machine Learning, 50 (1-2), 2003, pp. 95–125.
12. (Gal & Gahramani, 2016) Yarin Gal and Zoubin Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in Proc. 33rd International Conference on Machine Learning (ICML-16), 2016.
13. (Gelman & Rubin, 1992) A. Gelman, D. B. Rubin, "Inference from iterative simulation using multiple sequences", Statistical Science, 7, 1992, pp. 457–472.
14. (Geng & Hamilton, 2006) L. Geng, H. J. Hamilton, "Interestingness measures for data mining: a survey", ACM Computing Surveys, 38, no. 3, 2006, article 9.
15. (Ghoshal et. al, 2020) B. Ghoshal, C. Lindskog, A. Tucker, M. R. Berthold et al. Estimating Uncertainty in Deep Learningfor Reporting Confidence: An Applicationon Cell Type Prediction in Testes Basedon Proteomics (Eds.): IDA 2020, LNCS 12080, pp. 223–234, 2020.
16. (Gómez-Rubio, 2020) V. Gómez-Rubio, Bayesian inference with INLA, CRC Press, 2020.
17. (Grzegorczyk & Husmeier, 2008) M. Grzegorczyk, D. Husmeier, "Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move", Machine Learning, 71 (2-3), 2008, pp. 265.
18. (Hartnack et al., 2019) S. Hartnack, T. Odoch, G. Kratzer, R. Furrer, Y. Wasteson, T. M. L'Abée-Lund, E. Skjerve, "Additive Bayesian networks for antimicrobial resistance and potential risk factors in non-typhoidal Salmonella isolates from layer hens in Uganda", BMC Veterinary Research, 15, 212, 2019, pp. 1-9.
19. (Heider, 2013) D. Heider, R. Senge, W. Cheng, E. Hullermeier. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance and prediction. In: Bioinformatics, 2013, pp. 29(16):1946-52

20. (Held & Bove, 2014)L. Held, D. Sabanes Bove, Applied statistical inference, Springer, 2014.
21. (Herrera et al., 2016) F. Herrera, F. Charte, A. J. Rivera, M. J. del Jesus. Multilabel Classification Problem Analysis, Metrics and Techniques. In: Springer Verlag, 2016
22. (Imoto et al., 2003) S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks", Proc IEEE Comput. Soc. Bioinform. Conf. 2003, 2, pp. 104-113
23. (Jensen, 2001) F. V. Jensen, Bayesian network and decision graphs, Springer-Verlag, New York, 2001.
24. (Kratzer & Furrer, 2019) G. Kratzer, R. Furrer, mcmcabn: a structural MCMC sampler for DAGs learned from observed systemic datasets. R package version 0.3, 2019, https://CRAN.R-project.org/package=mcmcabn.
25. (Kratzer et al., 2020) G. Kratzer, F. I. Lewis, B. Willi, M. L. Meli, F. S. Boretti, R. Hofmann-Lehmann, P. Torgerson, R. Furrer, S. Hartnack, "Bayesian network modeling applied to feline calivicius infection among cats in Switzerland", Frontiers in Veterinary Science, 7, 2020, pp. 1-16.
26. (Kratzer et al., 2019) G. Kratzer, M. Pittavino, F. I. Lewis, R. Furrer, abn: an R package for modelling multivariate data using additive Bayesian networks. R package version 2.2, 2019, https://CRAN.R-project.org/package=abn.
27. (Koivisto & Sood, 2004) M. Koivisto, K. Sood, "Exact Bayesian structure discovery in Bayesian networks", Journal of Machine Learning Research, 5, 2004, pp. 549–573.
28. (González et al., 2019) L. I. Lopera González, A. Derungs, O. Amft, "A Bayesian approach to rule mining", 2019, arXiv:1912.06432.
29. (Lakshminarayanan et al., 2017) B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in Proc. Conference on Neural Information Processing Systems (NIPS), 2017.
30. (Ludwig et al., 2013) A. Ludwig, P. Berthiaume, P. Boerlin, S. Gow, D. Léger, F. I. Lewis, "Identifying associations in Escherichia coli antimicrobial resistance patterns using additive Bayesian networks", Preventive Veterinary Medicine, 110, 2013, pp. 64-75.
31. (MacKay, 2003) D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.
32. (Madigan & York, 1995) D. Madigan, J. York, "Bayesian graphical models for discrete data", International Statistical Review, 63, 1995, pp. 215–232.
33. (Mureşan et al., 2020) R. Mureşan, C. Zaharia, D. Zaharie, "Using Additive Bayesian Networks and Association Rules in Antimicrobial Resistance Analysis", In: 2020 E-Health and Bioengineering Conference (EHB). IEEE, 2020, pp. 1-4.
34. (NARMS) National Antimicrobial Resistance Monitoring System (NARMS) Now: Human Data. Atlanta, Georgia: U.S. Department of Health and Human Services, CDC. 09/18/2019. https://www.cdc.gov/narmsnow Accessed 7/19/2019.
35. (NCBI) NCBI Antimicrobial Resistance Resources https://www.ncbi.nlm.nih.gov/ pathogens/ antimicrobial-resistance/resources/
36. (Read, 2009) J. Read, B. Pfahringer, et al. Classifier Chains for Multi-label Classification. In: W. Buntine, M. Grobelnik, D. Mladenic, J. Shawne-Taylor (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD, LNCS 5782, 2009.
37. (Riemenschneider, 2016) Riemenschneider, M., Senge, R., Neumann, U. *et al.* Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *BioData Mining* **9,** 10 (2016). https://doi.org/10.1186/s13040-016-0089-1
38. (Rissanen, 1987)J. Rissanen, "Stochastic complexity", Journal of the Royal Statistical Society, Series B, 49 (3), 1987, pp. 223–229.
39. (Su & Borsuk, 2016) C. Su, M. E. Borsuk, "Improving structure MCMC for Bayesian networks through Markov blanket resampling", The Journal of Machine Learning Research, 17 (1), 2016, pp. 4042-4061.

40. (Tadesse et al., 2012) D. A. Tadesse, S. Zhao, E. Tong, S. Ayers, A. Singh, M. J. Bartholomew et.al., "Antimicrobial drug resistance in Escherichia coli from humans and food animals", United States,1950–2002, Emerging Infectious Diseases, 18 (5), 2012, pp. 741-749.
41. (Tan et al., 2004) P. -N. Tan, V. Kumar, J. Srivastava, "Selecting the right objective measure for association analysis", Information Systems, 29, 2004, pp. 293-313.
42. (Vladu & Zaharie, 2020)C. Vladu, D. Zaharie, Multi-label classification methods used in antimicrobial resistance prediction, Conferinţa Scolilor Doctorale din Consorţiul Universitaria, https://profs.info.uaic.ro/~CSDCU_MIF2020/wp-content/uploads/2020/10/64.pdf
43. (Werhli & Husmeier, 2007) A. V. Werhli, D. Husmeier, "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge", Statistical Applications in Genetics and Molecular Biology, 6, 2007, article 15, pp. 1-45.
44. (Wolpert & Macready, 1997) D. H. Wolpert, W. G. Macready, "No free lunch theorems for optimization, IEEE transactions on evolutionary computation", 1 (1), 1997, pp. 67-82.
45. (Zaharia et al., 2019) C. Zaharia, R. Mureşan, R. Moleriu, D. Zaharie, "Analysis of association measures used to discover antimicrobial resistance patterns", In: 2019 E-Health and Bioengineering Conference (EHB). IEEE, 2019, pp. 1-4.
46. (Zhang, 2007) M.L. Zhang, Z.H. Zho. ML-kNN: A Lazy Learning Approach to Multi-Label Learning. In: Pattern Recognition, 2007, pp:40(7), 2038-2048