Cuprins

[Introducere 3](#_Toc9281698)

[1 Descrierea problemei 4](#_Toc9281699)

[1.1 Căutare pe șiruri 4](#_Toc9281700)

[1.2 Descompunerea QR 4](#_Toc9281701)

[2 Resurse utilizate 5](#_Toc9281702)

[2.1 Subcapitol 1 5](#_Toc9281703)

[2.2 Subcapitol 2 5](#_Toc9281704)

[3 Rezolvarea problemei 6](#_Toc9281705)

[3.1 Metoda exhaustivă 6](#_Toc9281706)

[3.2 Metoda eficientă 6](#_Toc9281707)

[4 Abordarea problemei utilizând metode euristice 7](#_Toc9281708)

[4.1 Structură 7](#_Toc9281709)

[4.2 Reprezentarea datelor 7](#_Toc9281710)

[4.3 Operatori 8](#_Toc9281711)

[4.3.1 *Crossover* 8](#_Toc9281712)

[4.3.2 *Mutation* 11](#_Toc9281713)

[4.4 Funcția *fitness* 11](#_Toc9281714)

[4.5 Condiții de oprire 11](#_Toc9281715)

[5 Algoritmi euristici 12](#_Toc9281716)

[5.1 Algoritm genetic naiv 12](#_Toc9281717)

[5.1.1 *Tournament selection* 12](#_Toc9281718)

[5.1.2 *Roulette Wheel selection* 12](#_Toc9281719)

[5.2 *Hill* *climbing* 12](#_Toc9281720)

[5.3 *Simulated* *annealing* 13](#_Toc9281721)

[5.4 *Building* *blocks* 13](#_Toc9281722)

[5.5 Algoritmi cu restricție 13](#_Toc9281723)

[6 Euristici în combinație cu metoda eficientă 14](#_Toc9281724)

[6.1 Integrare 14](#_Toc9281725)

[6.2 Performanțe 14](#_Toc9281726)

[7 Implementare 15](#_Toc9281727)

[7.1 Structura generală 15](#_Toc9281728)

[7.2 Fișier de configurație 15](#_Toc9281729)

[7.3 Funcții de bază 15](#_Toc9281730)

[8 Analiza outputurilor 16](#_Toc9281731)

[8.1 Rezultate *VS* timp 16](#_Toc9281732)

[8.2 Avantaje 16](#_Toc9281733)

[8.3 Dezavantaje 16](#_Toc9281734)

[8.4 Performanțe 16](#_Toc9281735)

# Introducere

Motivație –

Gradul de noutate –

Obiective-

Structură , titlul capitolelor și legatura dintre ele

# Descrierea problemei

Text descriere

## Căutare pe șiruri

Text

## Descompunerea QR

Text

# Resurse utilizate

Text descriere

## Subcapitol 1

Text

## Subcapitol 2

Text

# Rezolvarea problemei

Text descriere

## Metoda exhaustivă

Utilizarea elementelor de combinatorica pentru a analiza tot setul de date

## Metoda eficientă

Utilizare descompunerii QR intr-un mod eficient, astfel incat aceasta nu se va calcula pentru fiecare sub-model (explicatie)

# Abordarea problemei utilizând metode euristice

Am ales această abordare prin metode euristice în primul rând pentru că problema are un spațiu de căutare foarte vast, numărul de sub-modele crescând exponențial odată cu creșterea numărului de coeficienți, astfel crescând și timpul de execuție al programului. În al doilea rând pentru a studia și compara diferite tipuri de implementări și abordări și nu în ultimul rând pentru că aceste metode euristice sunt de actualitate, astfel lăsând loc de idei originale.

În acest capitol voi introduce elementele cheie utilizate în algoritmii euristici, voi aminti mai multe tipuri, însa din acestea le voi discuta doar pe cele utilizate în aplicație împreună cu o scurtă motivație pentru care le-am ales pe unele în defavoarea celorlalte.

## Structură

Fiind metode euristice, nu tot setul de date este evaluat, iar structura la baza unui algoritm euristic pornește de la generarea unui punct de plecare *random*. Acesta poate fi un singur individ, sub-model sau o întreagă populație de sub-modele. Un sub-model reprezintă un posibil candidat al celei mai bune soluții din spațiul de căutare.

Pentru a decide dacă un sub-model este bun se calculează o funcție de cost denumită funcție *fitness* care asociază fiecărui sub-model o valoare numerică. Această funcție poate fi minimizată sau maximizată în funcție de problema pe care se aplică un astfel de algoritm.

Trecerea de la o iterație la alta, sau cu alte cuvinte, plasarea punctului de plecare mai departe în spațiul de căutare, se face cu ajutorul operatorilor despre care voi discuta în subcapitolul 4.3. Algoritmul poate fi oprit alegând diverse metode de oprire pe care le voi descrie la finalul acestui capitol.

## Reprezentarea datelor

În algoritmii euristici cea mai des întâlnită reprezentare a datelor este reprezentarea binară, *array*-uri de biți. Deoarece setul de date utilizat are în componență o matrice de linii și coloane, fiecare coloană reprezentând o caracteristică diferită, un element de compoziție, am ales aceeași reprezentare binară. Un sub-model reprezintă o combinație a coloanelor matricei. Astfel, un sub-model va reprezenta în prealabil un șir de biți de 0 și 1, peste care am adăugat o altă reprezentare și anume, pentru fiecare bit setat pe 1, indexul acestui bit va reprezenta exact indexul coloanei alese pentru a alcătui sub-modelul.

De exemplu : având o matrice de 6 coloane numerotate de la 0 la 5, șirul 010110 este reprezentarea sub-modelului format din matricea de coloane : 1,3 și 4.

De asemenea, în algoritmii implementați, am utilizat limbajul algoritmilor genetici și anume:

* Un sub-model reprezintă o combinație de coloane, dar este utilizat și sub numele de individ, *genotype* sau *chromosome*
* Un individ este alcătuit dintr-un șir de biți setați pe 1 sau 0, acești biți mai sunt utilizați sub numele de gene

## Operatori

Operatorii sunt algoritmi aplicați peste sub-modele cu scopul de a genera alte sub-modele noi, diferite , însa cu premisa că noile sub-modele păstrează un anumit procent din material genetic al sub-modelului părinte. Mai precis, sub-modelul nou obținut va păstra șiruri de biți de 0 și 1 neschimbate. În cele mai multe cazuri acești operatori transformă sub-modelul prin procedee *random*.

### *Crossover*

Un operator des utilizat este operatorul de *crossover*. Acesta este aplicat pe doi indivizi denumiți părinți și generează alți doi indivizi sub numele de copii, care vor avea o combinație a materialului genetic a celor doi indivizi părinte. Poate fi de mai multe tipuri (*1-point crossover, n-point crossover, uniform crossover, flat crossover,* etc) însă cei utilizați în această lucrare sunt:

* *1-point crossover*
* *1-point crossover* adaptat
* *uniform crossover*
* *RRC[[1]](#footnote-1) crossover*

Metoda *1-point crossover* este cea mai simplă și comună operație de *crossover*. Se alege aleator un punct de taiere al indivizilor, acesta va fi comun atât pentru primul individ părinte cât și pentru cel de-al doilea. Primul individ copil va moșteni șirul de biți de la primul individ părinte până la punctul de tăiere, la care, se adaugă șirul de biți moștenit de la cel de-al doilea individ părinte, după punctul de tăiere. Cel de-al doilea individ copil va moșteni șirurile de biți rămase de la indivizii părinte, cele care nu au fost selectate pentru primul individ copil.

De exemplu, pentru punctul de tăiere de valoare 1 rezultate sunt:

* Părinte1 : 10|0100
* Părinte2 : 00|1011
* Copil1 : 10|1011
* Copil2 : 00|0100

Metoda *1-point crossover* adaptat este o modificare a metodei *1-point crossover*. La fel ca și la metoda de bază, se alege aleator un punct de tăiere al indivizilor, acesta va fi comun atât pentru primul individ părinte cât și pentru cel de-al doilea. Diferența este că acest punct de tăiere va respecta o condiție, și anume acesta este ales în așa măsură încât prin aplicarea metodei de *1-point crossover* adaptat, indivizii copii vor păstra același număr de gene de valoare 1 ca și indivizii părinte. Această condiție impune ca înainte de punctul de taiere cât și după acesta, numărul de gene cu valoarea 1 trebuie sa fie egal, condiție ce trebuie îndeplinită simultan de ambii indivizi părinte. Această metodă este creată de mine și am utilizat-o în Algoritmi cu restricție pe care îi voi descrie în subcapitolul 5.5.

Metoda *uniform crossover* este o metodă care nu aduce foarte multe schimbări celor doi indivizi copil, daca indivizi părinte au gene asemănătoare. La această metodă pentru fiecare dintre genele de pe aceeași poziție din indivizi părinte care sunt diferite, se va alege cu o probabilitate generată aleator ce decide dacă indivizii părinte vor face *switch* între aceste gene. Indivizii copil vor fi apoi fiecare o copie a noilor indivizi părinte. Condiția ca doua gene sa fie interschimbate este ca probabilitatea generată aleator sa fie mai mare decât o probabilitate de interschimbare stabilită în prealabil.

De exemplu, pentru o probabilitate de *switch* între indivizii părinte de 50% în care pentru fiecare din genele diferite s-a generat o probabilitate aleatorie mai mare de 50% rezultatele sunt:

* Părinte1 : 100100
* Părinte2 : 001011
* Copil1 : 001011
* Copil2 : 100100

Un alt exemplu pentru o probabilitate de *switch* între indivizii părinte de 50% în care pentru ultimele două gene diferite s-a generat o probabilitate mai mică de 50% rezultatele sunt:

* Părinte1 : 100100
* Părinte2 : 001011
* Copil1 : 001000
* Copil2 : 100111

Metoda *RRC crossover* este o metodă foarte asemănătoare celei descrise anterior, *uniform crossover.* În același fel, se generează un număr aleator pentru acele gene care diferă pentru indivizii părinte. De această dată însă, se generează o probabilitate aleatoare pentru fiecare individ copil, și nu se va face o interschimbare a acestor gene, ci în funcție de probabilitatea generată, dacă aceasta este mai mare decât probabilitatea stabilită în prealabil, va indica dacă pe acea poziție individul copil va prelua gena de la primul individ părinte sau de la cel de-al doilea.

De exemplu, pentru o probabilitate de preluare a genelor indivizilor părinte de 50% în care pentru fiecare din genele diferite s-a generat o probabiliate1 a primului individ-copil aleatoriu mai mare de 50% și o probabiliate2 a celui de-al doilea individ-copil mai mică de 50% rezultatele sunt:

* Părinte1 : 100100
* Părinte2 : 001011
* Copil1 : 100100
* Copil2 : 001011

Un alt exemplu pentru o probabilitate de preluare a genelor indivizilor părinte de 50% în care pentru fiecare din genele diferite s-a generat o probabiliate1 a primului individ copil aleatoriu mai mare de 50% și o probabiliate2 a celui de-al doilea individ copil tot mai mare de 50%, însă pentru ultimele 2 gene probabiliatea2 este mai mica de 50% rezultatele sunt:

* Părinte1 : 100100
* Părinte2 : 001011
* Copil1 : 100100
* Copil2 : 100111

Ca o mică concluzie, această ultimă metodă descrisă duce la generarea unor noi indivizi care păstrează tot mai mult structura materialului genetic ale celor doi indivizi părinte. Este o metodă care oferă rezultate mai puțin aleatoare. Din acest motiv, când generațiile devin din ce în ce mai stabile și mai evoluate, acele gene care duc la un *fitness* mai bun, vor avea o șansă mai mică de a se pierde. Această metodă de *crossover* este foarte utilă de exemplu în algoritmul genetic bazat pe *Building blocks* pe care îl voi discuta în subcapitolul 5.4 . Cu toate acestea și metodele bazate mai mult pe combinarea în mod aleator și generarea unor indivizi copil foarte diferiți de indivizii părinte pot avea un impact major în algoritmii care se opresc într-un minim local deoarece pot duce la modificare *landscape*-ului de căutare, metode cu rezultate foarte bune în algoritmul *Simulated annealing* pe care îl voi discuta în subcapitolul 5.3.

### *Mutation*

Un alt operator utilizat în obținerea noilor generații de indivizi este operatorul de mutație. O diferență majora între *crossover* și *mutation* este faptul că la generarea unui nou individ copil este nevoie de un singur individ părinte. O altă diferență o constituie faptul că cel mai des utilizate metode de mutație aduc modificări la un număr mic de gene. Ca și în cazul operatorului *crossover* există mai multe metode de mutație studiate (*insert muatation, inversion mutation, uniform mutation, etc*) însă cele utilizate de către mine sunt:

* *flip mutation*
* *interchanging mutation*
* *absolute interchanging mutation*
* *reversing mutation*

Metoda *flip mutation* este cea mai des utilizată metoda de către mine în această lucrare. Aceasta constă în generarea unui număr aleator de gene n ce constituie numărul de gene al individului părinte care vor aduce schimbări în individul copil. Așadar se generează n gene aleatoare care își vor schimba valoare curentă cu cea opusă. Prin urmare, dacă gena de pe una din cele n poziții aleatore generate are valoarea 0 în individul părinte, aceasta va lua valoare 1 în individul copil, în caz contrar daca gena are valoare 0, aceasta va lua valoarea 1, restul genelor sunt copiate așa cum sunt în individul copil.

De exemplu, dacă n are valoarea 3 și pozițiile generate au valorile 0, 2, și 3, atunci se obține:

* Părinte : 100100
* Copil : 001000

Metoda *interchanging mutation* este o metodă care nu aduce foarte multe schimbări individului copil. Se generează doua poziții aleatoare, iar genele de pe acele poziții fac schimb de valori în individul copil, restul genelor fiind copiate in individul copil așa cum sunt. Această metodă aduce avantajul păstrării numărului de gene de valoare 1 din individul părinte în individul copil nou creat. Mai concret, păstrează același conținut, modificând doar ordinea genelor. Este o metodă pe care am utilizat-o în Algoritmi cu restricție despre care voi vorbi în subcapitolul 5.5.

De exemplu, dacă se generează aleatoriu pozițiile 0 și 5, atunci se obține:

* Părinte : 100100
* Copil : 000101

Metoda *absolute interchanging mutation* este o metodă adaptată de mine după modelul celei prezentate anterior. Această metodă urmează aceeași procedura de generare a două poziții aleatoare pentru a fi interschimbate valorile genelor cu două restricții. Una dintre ele este ca cele doua poziții *random* generate să fie diferite, iar cealaltă restricție impune ca cele doua gene de pe pozițiile generate să aibă valori diferite. Aceste condiții au fost introduse de mine pentru a genera de fiecare dată indivizi copii cu *genotype* diferit de cel al indivizilor părinte.

Metoda *reversing mutation* este o metodă care face excepție de la ideea că operatorul de mutație nu aduce foarte multe schimbări. Această metoda generează o poziție aleatoare astfel că până la aceasta poziție individul copil copie fiecare genă din individul părinte așa cum este, iar pentru genele ce urmează după această poziție, în individul copil se copie valoarea opusă a genelor din individul părinte. Daca poziția generată are valoarea 0, atunci individul copil este exact opusul individului părinte.

De exemplu, dacă se generează aleatoriu poziția 2 atunci se obține:

* Părinte : 100100
* Copil : 100011

Pentru obținerea unor rezultate cât mai bune, am introdus operatori care să aibă efect mai mare in generarea noilor populații de indivizi. Cel mai des am utilizat in combinație operatorul *RRC crossover* și *flip muation*.

## Funcția *fitness*

În problemele euristice este des utilizată funcția de *fitness* sau cu alte cuvinte funcția de cost. Această funcție este o reprezentare numerică ce exprimă la ce nivel de adaptare se clasifică un individ. Pe baza acestei funcții algoritmii execută căutarea celui mai bun candidat, astfel că joacă un rol foarte important în distincția unui individ adaptat față de unul mai puțin adaptat. Prin urmare, o funcție de cost are atribuția de a clasifica cât mai bine indivizii și de aceea aceasta poate să ajungă la un nivel de complexitate mare, ceea ce induce un timp de execuție depășit.

În rezolvarea găsirii celui mai bun candidat pe problema căutării celui mai bun set de coeficienți care să aproximeze soluția sistemului, discutat în capitolul 1, am utilizat funcția de RSS[[2]](#footnote-2). Această funcție calculează distanța euclidiană de la soluția sistemului la soluția aproximată, ceea ce înseamnă că, cu cât eroarea este mai mică cu atât soluția aproximată este mai bună. Prin urmare cu cât un candidat are funcția *fitness* cu o valoare mai mică, cu atât acest candidat este mai favorabil.

Dat fiind faptul că, pentru a calcula RSS-ul unui individ nu se ține cont și de dimensiunea acestuia, pentru a obține rezultate cât mai reale, am aplicat funcția de AIC[[3]](#footnote-3). Prin urmare, pe lângă obținerea mediei de eroare, am ținut cont pentru funcția de cost și de numărul de coloane selectate al candidatului. Această funcție AIC exprimă câtă informație din setul de date a fost luată în vederea obținerii funcției de cost. Cu cât eroarea este mai mică și tot odată informațiile pierdute mai puține, cu atât funcția va da un rezultat mai bun, în cazul de față în interesul meu este obținerea unei valori cât mai mici.

În concluzie, funcția *fitness* arată in felul următor:

AIC = n + n\*log 2π + n\*log(RSS/n) + 2\*(p + 1) ,

unde n semnifică numărul de coloane al setului de date, iar p înseamnă numărul de coloane al candidatului.

## Condiții de oprire

Întrucât algoritmii euristice nu verifică tot setul de date, aceștia necesită o condiție de oprire, alta decât atunci când nu mai există indivizi de verificat. Există multe metode studiate, unele mai eficiente decât altele, însă depind foarte mult de algoritm.

Voi preciza în acest subcapitol câteva dintre procedurile de oprire, însă le voi discuta mai amănunțit în capitolul ce urmează. Așadar, una din condițiile de oprire este atunci când algoritmul converge, ceea ce înseamnă că în iterațiile următoare algoritmul nu oferă un candidat mai adaptat față de cel deja clasificat ca cel mai bun. O altă condiție de oprire este atunci când numărul maxim de iterații stabilit în prealabil ajunge la valoarea maximă. De asemenea, când populația de indivizi nu este prea vastă, iar procedura de selecție a generației următoare devine prea strictă, populația poate să ajungă să fie vidă și atunci algoritmul este oprit. Un alt exemplu, strict pe algoritmul *Hill* *climbing* pe care îl voi discuta în capitolul următor, este atunci când procedura de a accepta un alt posibil candidat pentru a fi analizat devine imposibilă, întrucât temperatura se apropie de valoarea 0, în condițiile în care acest algoritm se bazează pe faptul că cu cât temperatura este mai mare cu atât probabilitatea unui candidat de a fi ales pentru generația următoare crește.

# Algoritmi euristici

Text descriere

## Algoritm genetic naiv

Algoritmii genetici fac parte din clasa algoritmilor euristici, dar mai mult decât atât, urmăresc modelul evoluției darwiniste. Au la bază o structură bine definită care însa poate fi parametrizată și modelată în favoarea setului de date.

O structură foarte simplă a algoritmului genetic pe care am abordat-o pentru a implementa algoritmul genetic naiv poate fi descrisă în câțiva pași. Pentru început am generat aleator o populație de indivizi, numărul de indivizi fiind configurabil în funcție de numărul de observații din setul de date. Am calculat *fitness*-ul tuturor indivizilor, salvând cel mai bun candidat. Pentru noua generație de indivizi ce urmează, am trecut indivizii printr-un proces de selecție, astfel nu toți indivizii vor produce generații noi de indivizi ci doar cei selectați. Voi descrie în următoarele două subcapitole metodele de selecție. După acest proces de filtrare al populației, am efectuat asupra indivizilor rămași, operații de *mutation* și *crossover* . În urma acestor operații genotipul indivizilor se modifică, mutând spațiul de căutare într-o nouă direcție. Înainte de a trece la următoare iterație a algoritmului, printr-un proces simplu de comparație a funcției *fitness* am actualizat individul cel mai adaptat, doar în cazul în care noua generație a produs un individ mai adaptat decât o generație anterioară. În fiecare iterație ce urmează, am efectuat aceeași pași și anume selecția populației, modificarea indivizilor prin intermediul operatorilor și salvarea celui mai bun individ, până ce algoritmul converge.

### *Tournament selection*

Metoda de selecție *Tournament selection* formează un *pool* temporar de dimensiune k pe care îl populează cu indivizi aleși în mod aleatoriu din populația curentă. Din acest *pool* alege pentru generația următoare pe cel mai bun candidat. Tot acest proces reprezintă un turneu de selecție. Cu cât procesul se repetă mai mult, cu atât generația următoare va avea dimensiune mai mare. Un aspect foarte important este alegerea acestui număr k, cu cât turneul trebuie să aleagă cel mai bun candidat dintr-un număr mare de candidați al acestui *pool* temporar, cu atât crește probabilitatea ca turneul să alegă același candidat pentru noua populație, ceea ce înseamnă că populația nu va fi diversificată. Acest lucru forțează algoritmul să fie blocat în același spațiu de căutare.

- implementare

### *Roulette wheel selection*

Metoda de selecție *Roullete wheel* urmează ideea că fiecare individ are șansa să fie ales pentru viitoarea generație chiar și în cazul în care funcția *fitness* îl clasifică ca fiind cel mai nefavorabil candidat. Astfel, fiecărui individ i se calculează probabilitatea de a fi selectat pe baza evaluării funcției *fitness* în raport cu suma *fitness*-ului tuturor indivizilor. Șansa de care vorbeam mai sus este dată de faptul că în momentul stabilirii dacă un individ este ales sau nu, am generat o probabilitate *random* , iar dacă aceasta este mai mare decât probabilitatea individului curent de a fi selectat și mai mică decât probabilitatea de selecție a următorului candidat, atunci individul este selectat.

- implementare

## *Hill* *climbing*

Text

## *Simulated* *annealing*

Text

## *Building* *blocks*

Text

## Algoritmi cu restricție

Text

# Euristici în combinație cu metoda eficientă

Text descriere

## Integrare

Text

## Performanțe

Text

# Implementare

Text

## Structura generală

## Fișier de configurație

## Funcții de bază

# Analiza outputurilor

Text descriere

## Rezultate *VS* timp

Text

## Avantaje

Text

## Dezavantaje

Text

## Performanțe

Text

1. RRC Este o abreviere a metodei *Random Respectful Crossover* [↑](#footnote-ref-1)
2. RSS – *Residual Sum of Squares* [↑](#footnote-ref-2)
3. AIC - *Akaike information criterion* [↑](#footnote-ref-3)