VÉGZŐS KONFERENCIA 2009
2009. május 20, Budapest

Újfajta, automatikus, döntési fa alapú adatbányászati módszer idősorok osztályozására

Hidasi Balázs

hidasi@tmit.bme.hu

Konzulens: Gáspár-Papanek Csaba

Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Távközlési és Médiainformatikai Tanszék

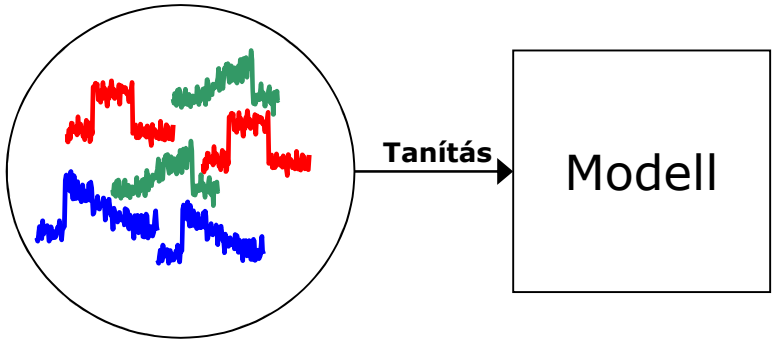
2009. Május 20.

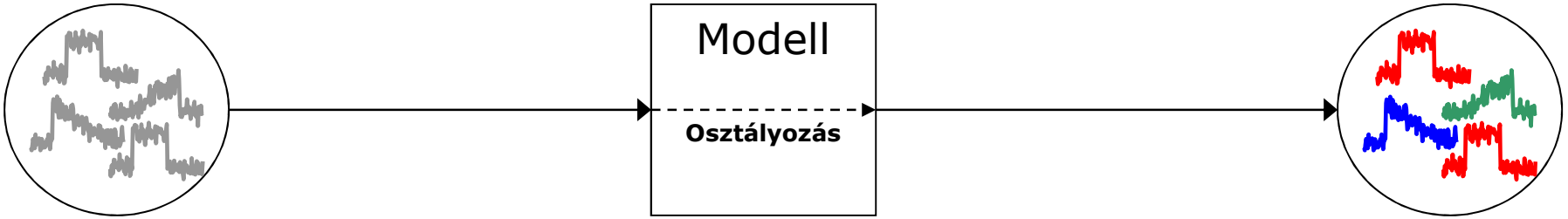
Tartalom

- Motiváció
- Célok
- A ShiftTree algoritmus
  - A módszer alapjai
  - Tanulás (ötlet)
  - Osztályozás példa
  - Optimalizálás: többszörös modellezés
- Eredmények
  - Benchmark
  - Verseny
- Alkalmazási lehetőségek
  - Beszélő felismerése
  - Gesztusfelismerés (+felhasználó azonosítás)
  - „Gondolatok” felismerése
- Összefoglalás

Motiváció

- Idősorok osztályozása
  - Sok az idősor-jellegű adat
  - Előtérbe kerülő alkalmazási területek
    - Nyomkövetés
    - Hangfelismerés
    - Diagnosztika
    - Gesztusfelismerés





- Jelenlegi algoritmusok hátrányai
  - Klasszikus módszerek
    - Jelentős emberi munka (előkészítés)
    - Nem erre találták ki
      - Információvesztés (pontatlanság)
    - Általában nem magyaráz
  - Terület-specifikus algoritmusok
    - Más területen nem hatékony

Célok

- Automatikus
  - Kevés emberi munka
    - Rövid előkészítési fázis
  - Minél több típus általános kezelése
    - Változók száma, osztályok száma, idősorok hossza, stb.
  - Több területen használható (általános)
- Pontos osztályozás
  - Magas találati arány
- Magyarázó
  - Könnyen értelmezhető modellt épít
  - Ellenőrizhető

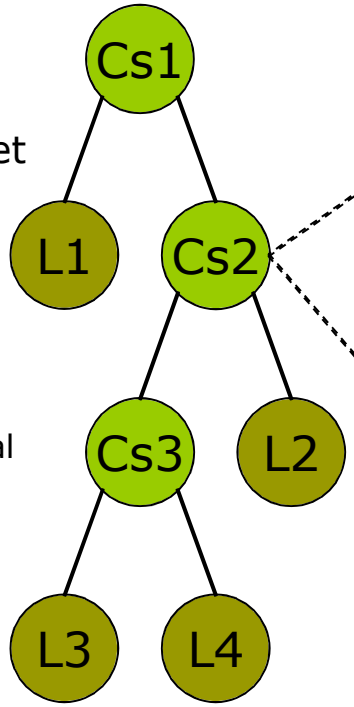
ShiftTree – A módszer alapjai

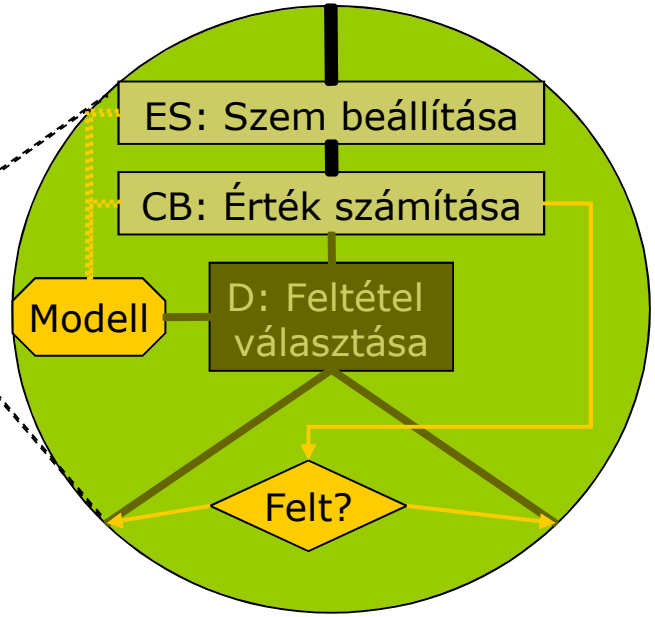
- Hibrid algoritmus
  - Döntési fa alap
    - Szerkezet
    - Vágások jóság értékei
    - Leállási feltételek
  - Módosított csomópont-szerkezet
- Moduláris felépítés

  - Szemtologató (EyeShifter)
    - ES-Operator (ESO)
    - Szem (pointer) mozgatás

  - Feltételállító (ConditionBuilder)
    - CB-Operator (CBO)
    - Érték származtatás a szem által mutatott értékből
      - (és környezetéből)

  - Döntő (Decider)
    - Vágási helyek vizsgálata
    - Jóságérték számítás
    - Optimális vágás választása a lehetségesekből
    - Feltétel kiszámítása





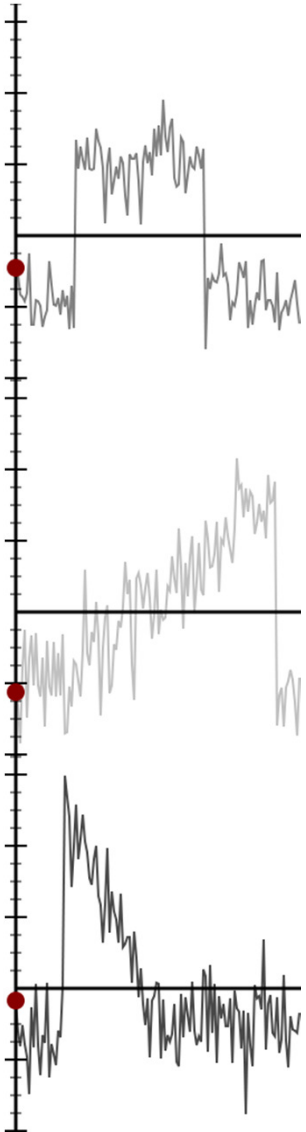
ShiftTree – Tanulás (ötlet)

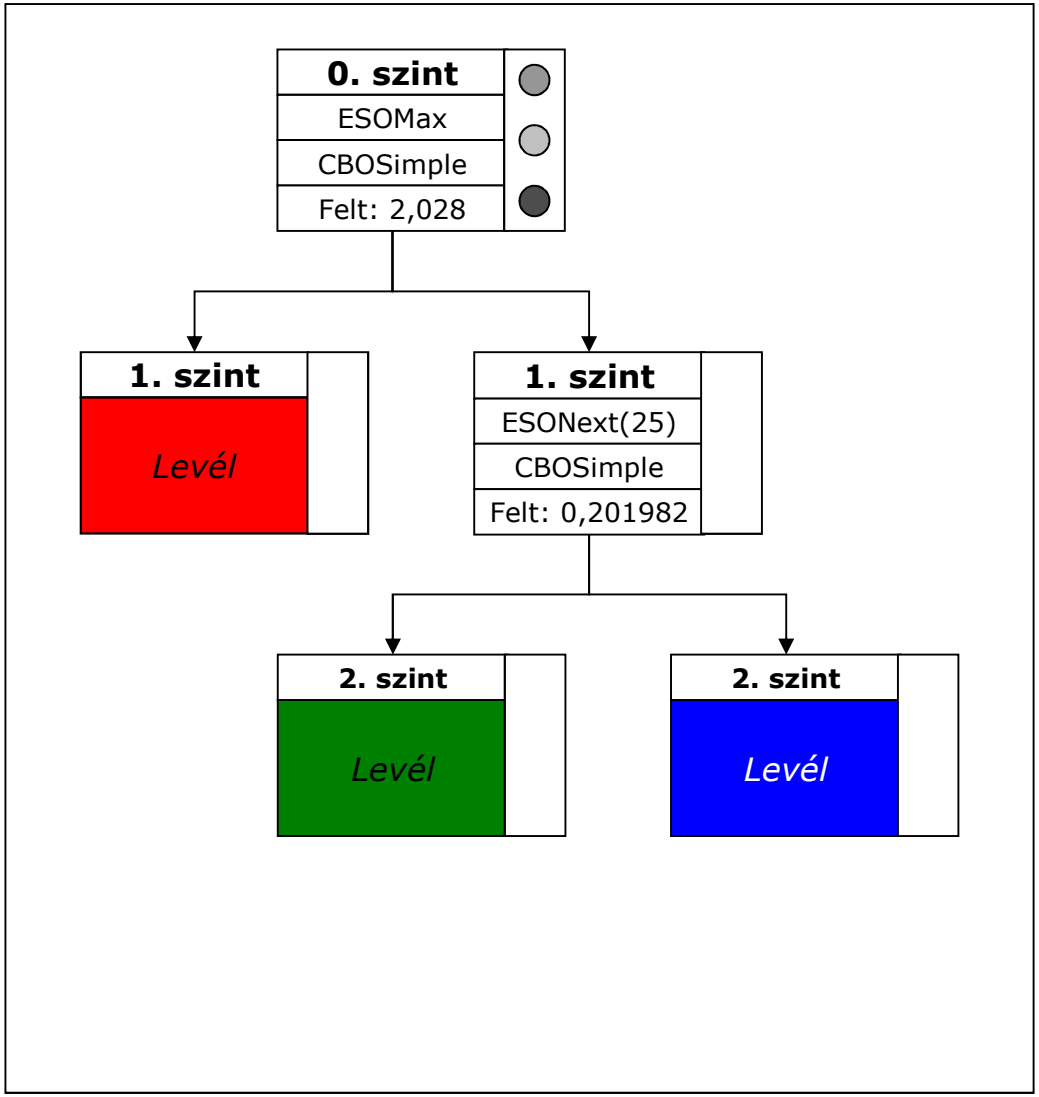
- „Dinamikus attribútumok”
  - Hol nézzük? (ESO)

    - 25 időegységgel előrefele (ESONext(25))
    - A globális maximumnál (ESOMax)
    - 60 méretű intervallumon belül a legkisebb értéknél
    - …
  - Mit nézzünk? (CBO)

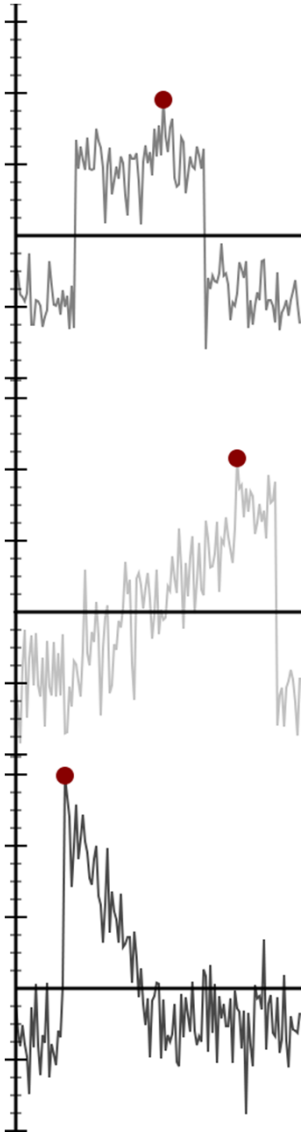
    - A pontbeli értéket (CBOSimple)
    - Az érték környezetének normális eloszlás szerinti súlyozott átlagát (CBONormal)
    - Az ugrás során a lokális maximumok számát
    - Az ugrás hosszát
    - …
- Tanulás egy csomópontban
  - Leállási feltételek vizsgálata
  - Lehetséges attribútumok kiszámolása (ESO-CBO párok)
  - Az (első) optimális vágás megtalálása (ezt végzi a Decider)
    - Attribútumok közül egy
    - Feltétel érték
  - Operátorok és a feltétel érték megjegyzése
  - Vágás az attribútum és a feltétel alapján
    - Kettéosztani a tanítópontokat a jobb és bal gyermeknek
  - Rekurzívan ugyanez a gyermek csomópontokban

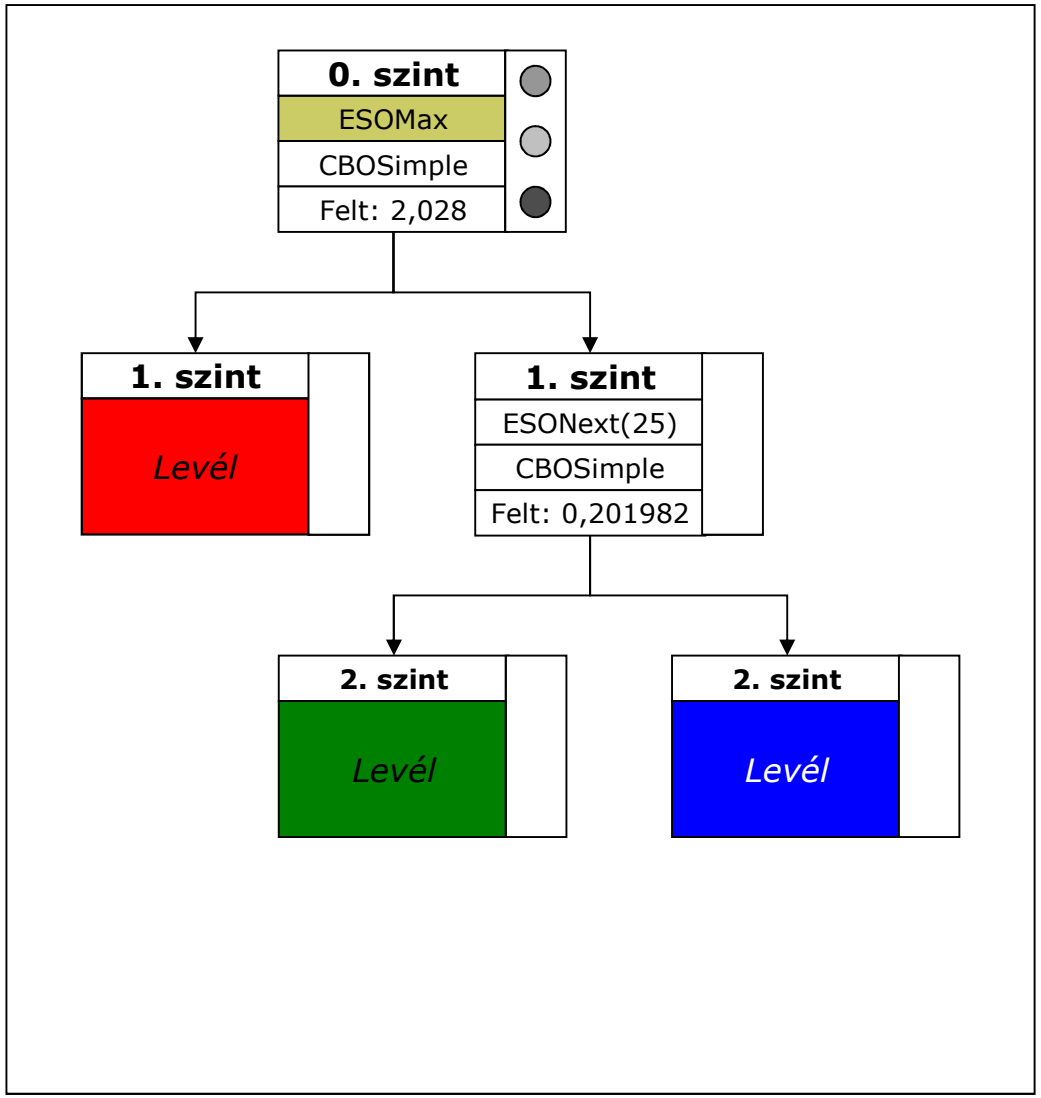
ShiftTree – Osztályozás példa



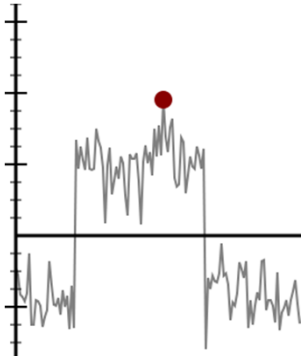


ShiftTree – Osztályozás példa

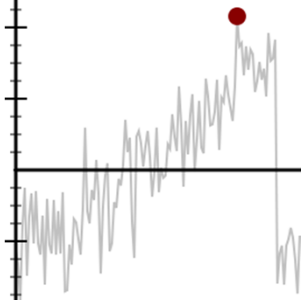




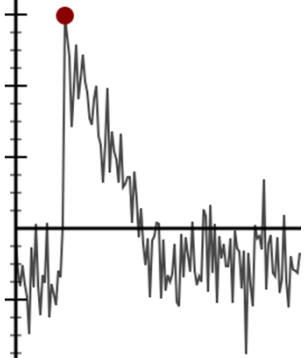
ShiftTree – Osztályozás példa



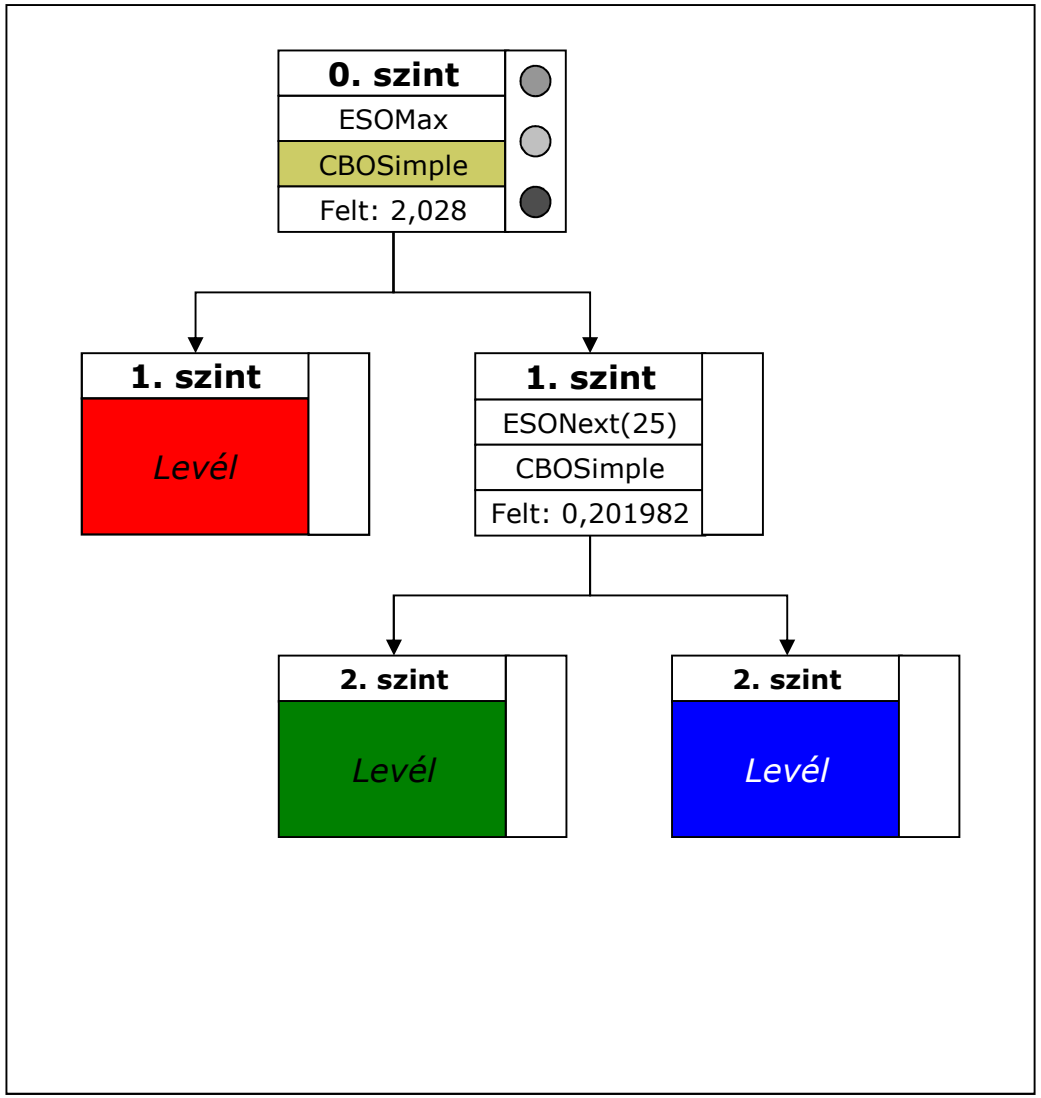
Érték:
1,89136



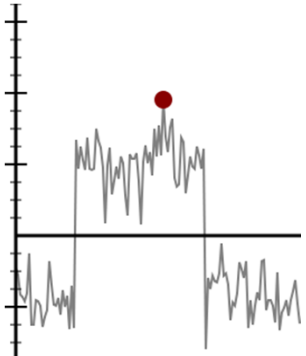
Érték:
2,15333



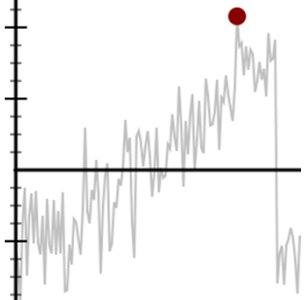
Érték:
2,97557



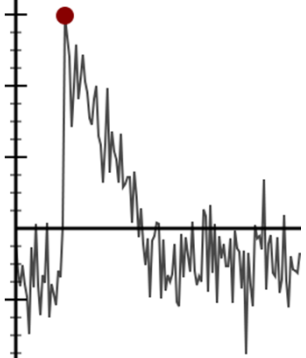
ShiftTree – Osztályozás példa



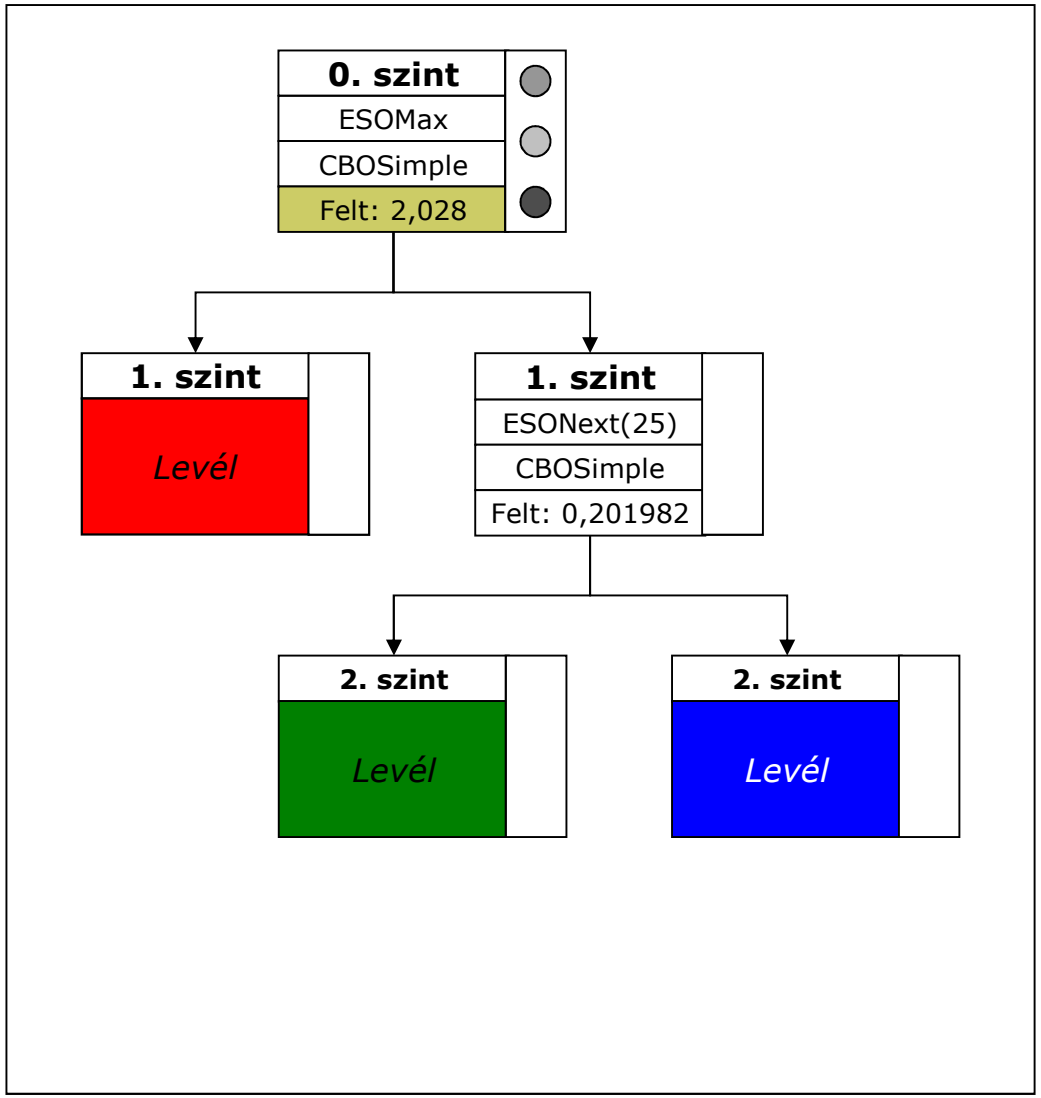
Érték:
1,89136



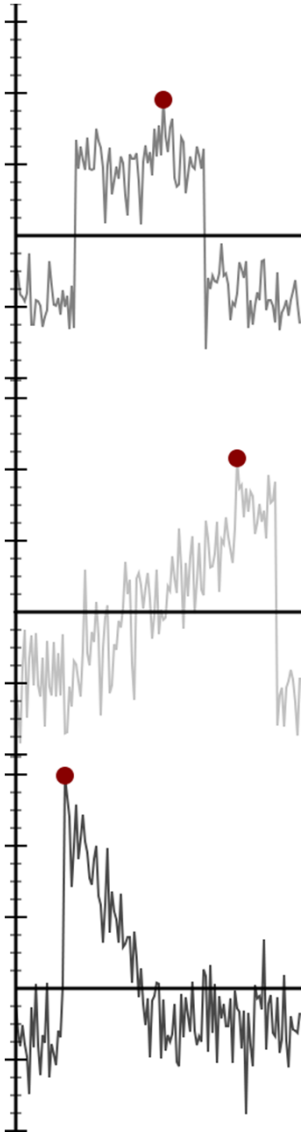
Érték:
2,15333

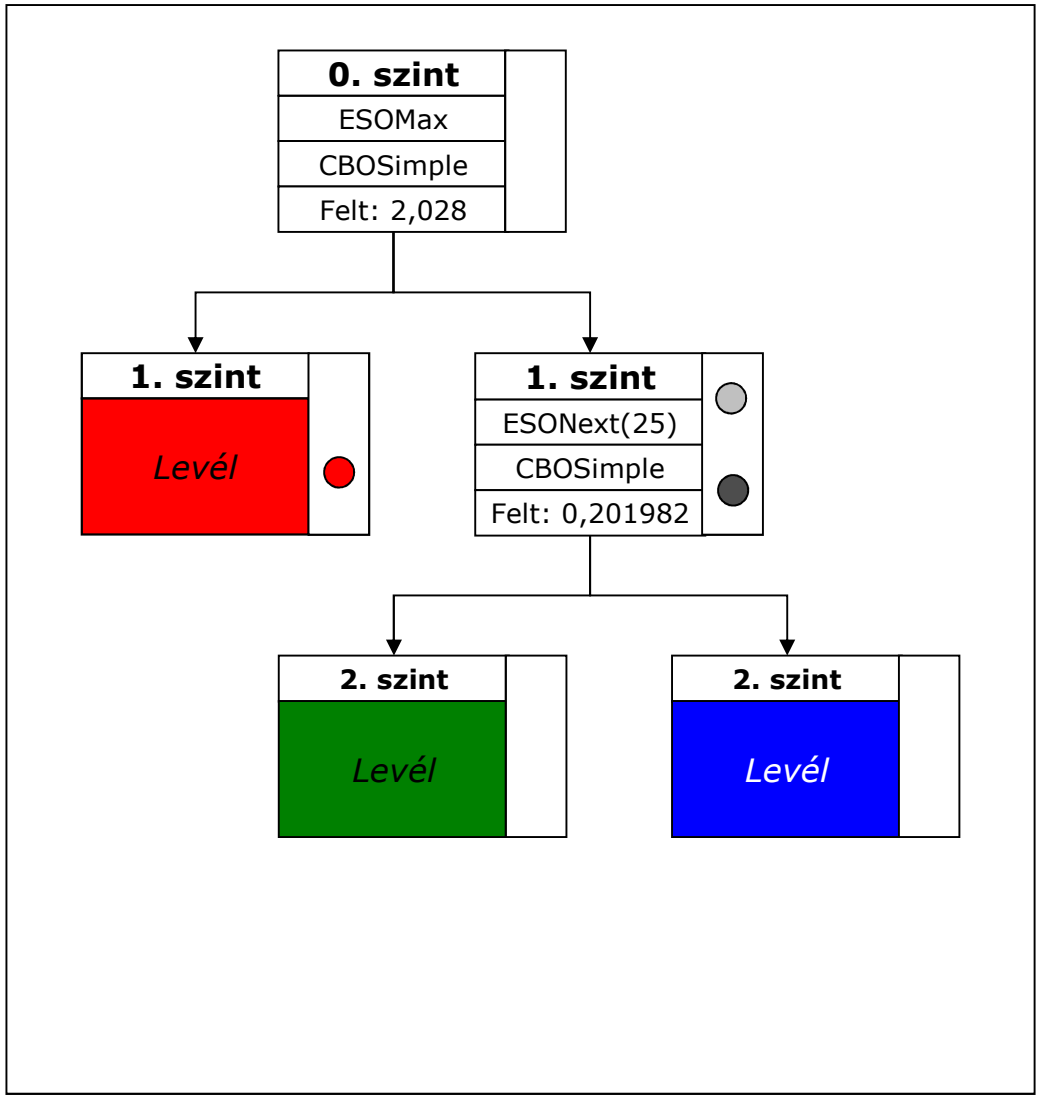


Érték:
2,97557

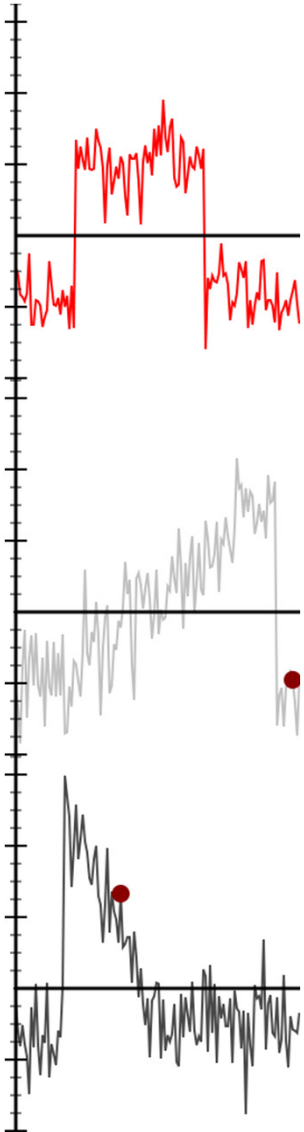


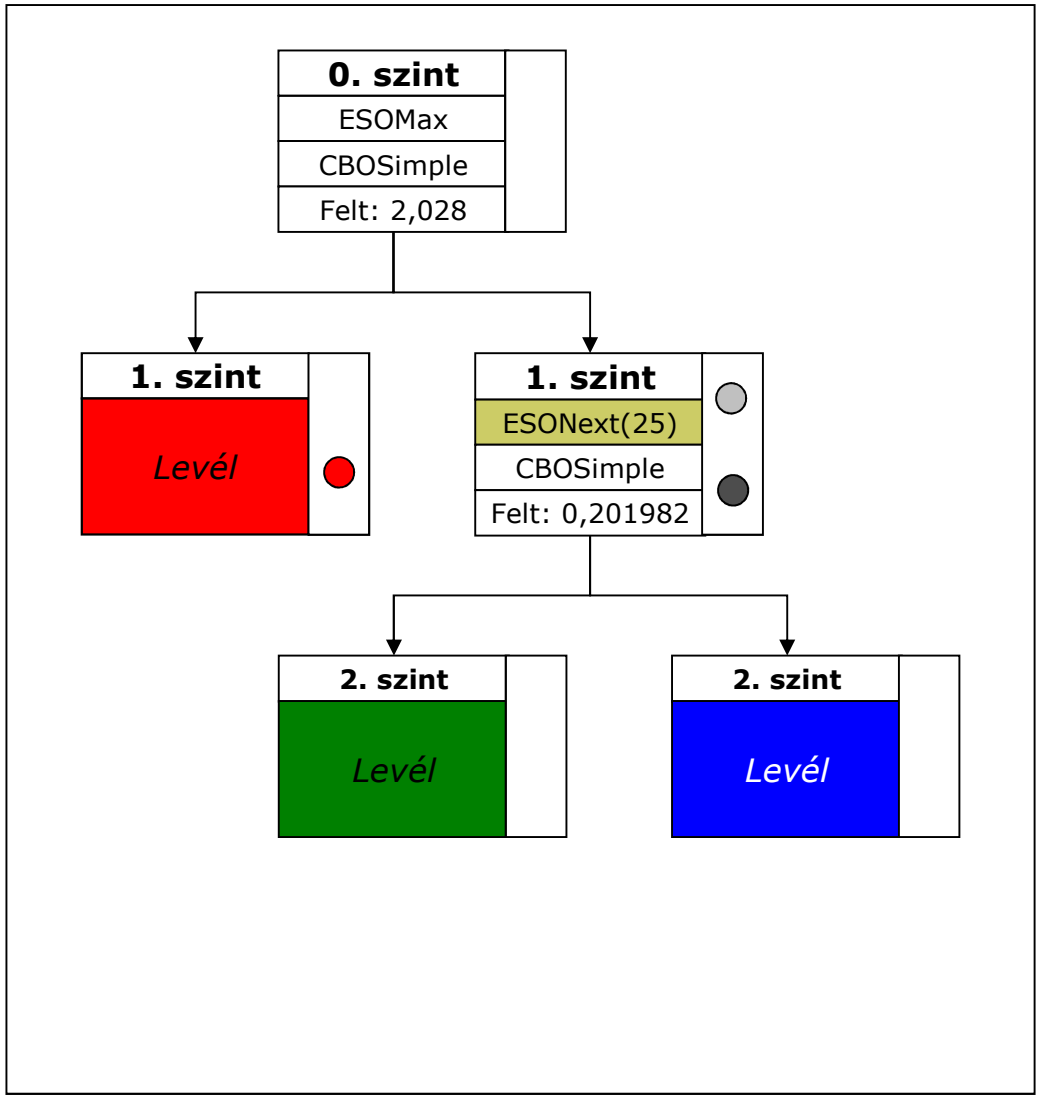
ShiftTree – Osztályozás példa



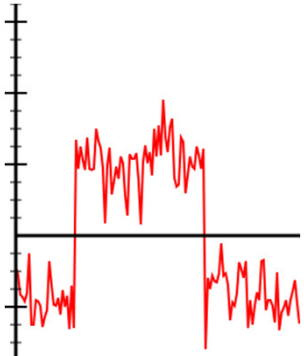


ShiftTree – Osztályozás példa

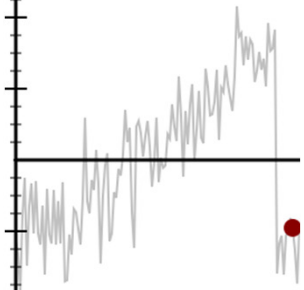




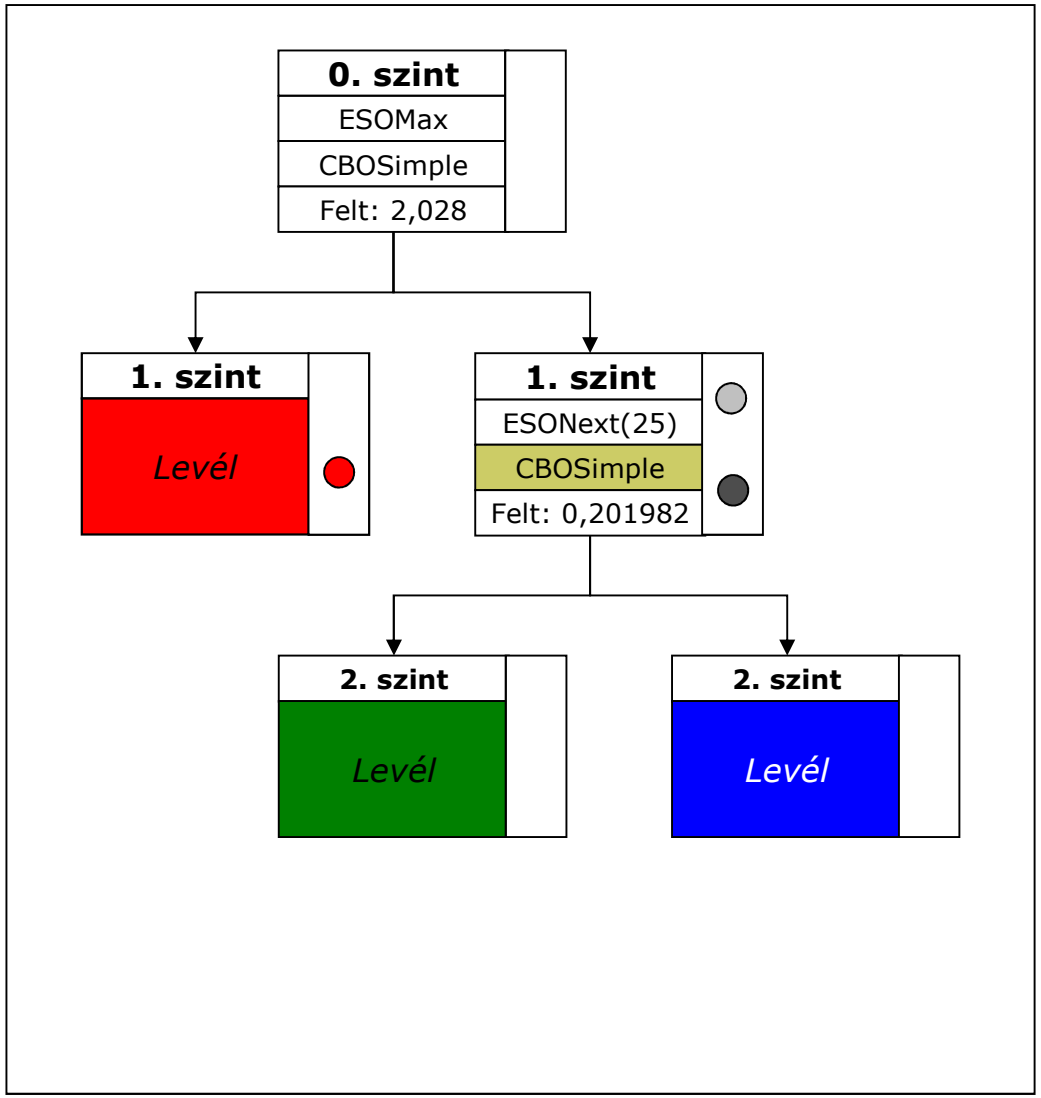
ShiftTree – Osztályozás példa



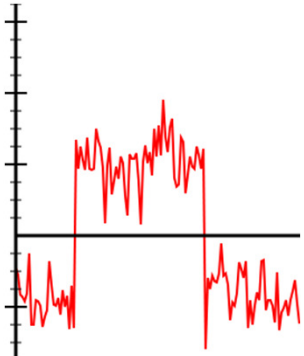
Érték:
-0,953538

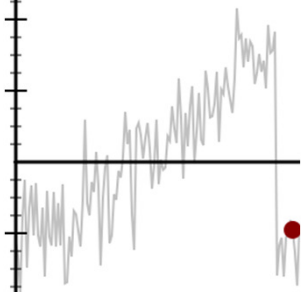


Érték:
1,32432

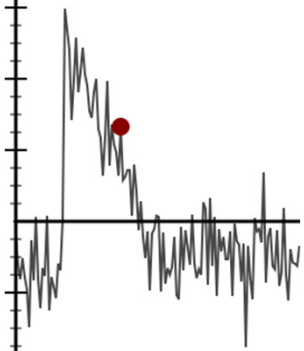


ShiftTree – Osztályozás példa

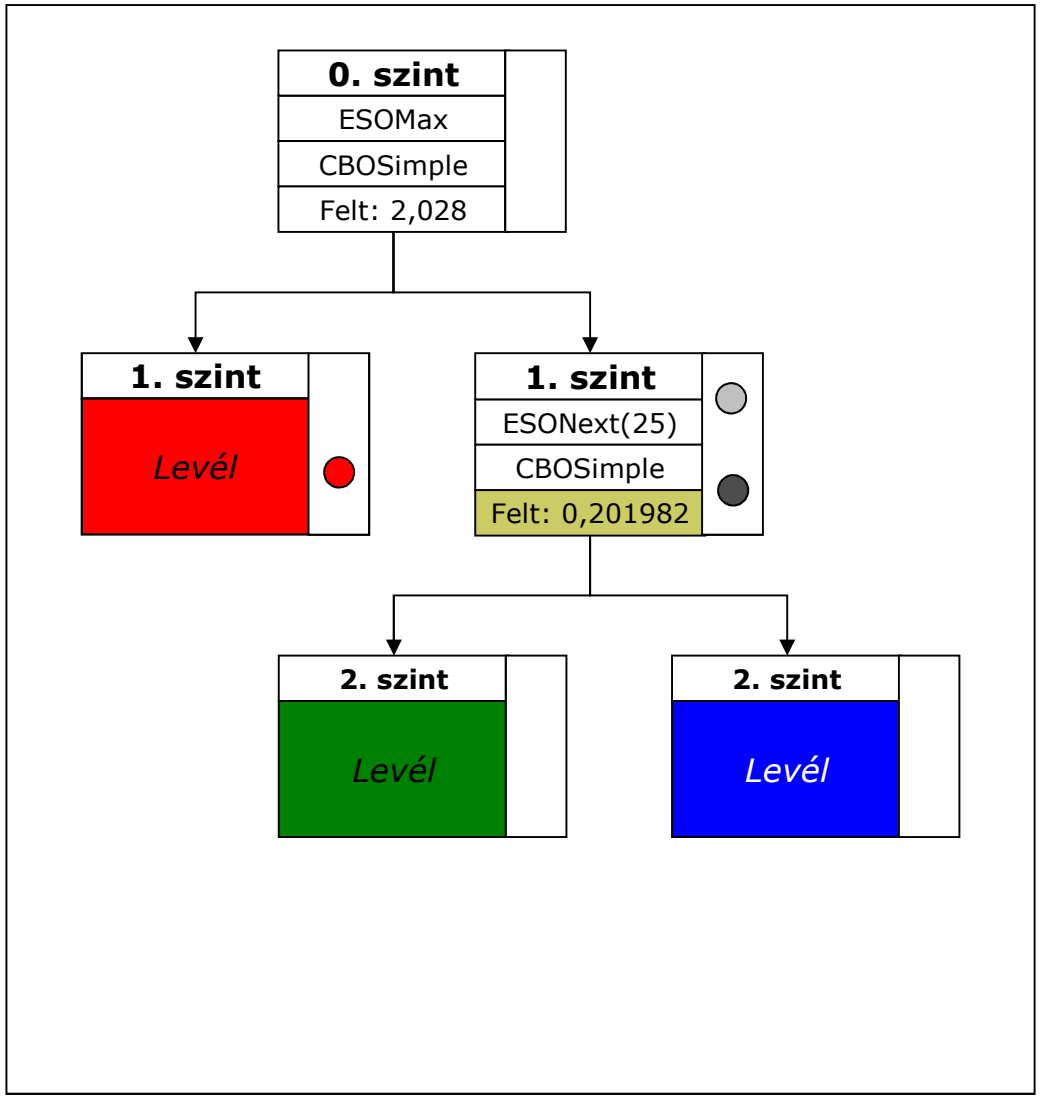




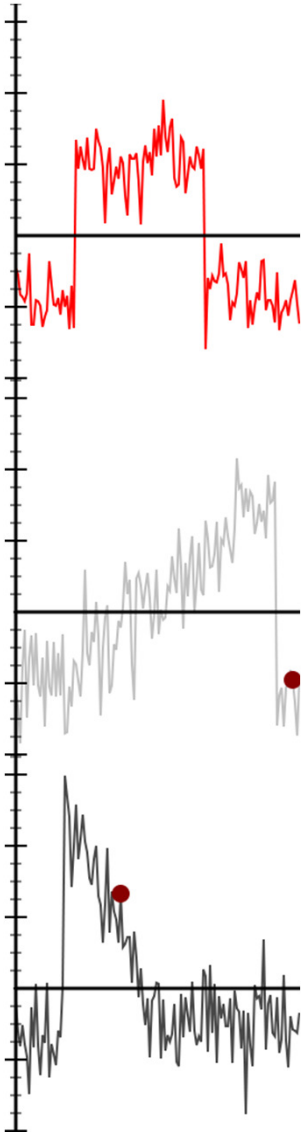
Érték:
-0,953538

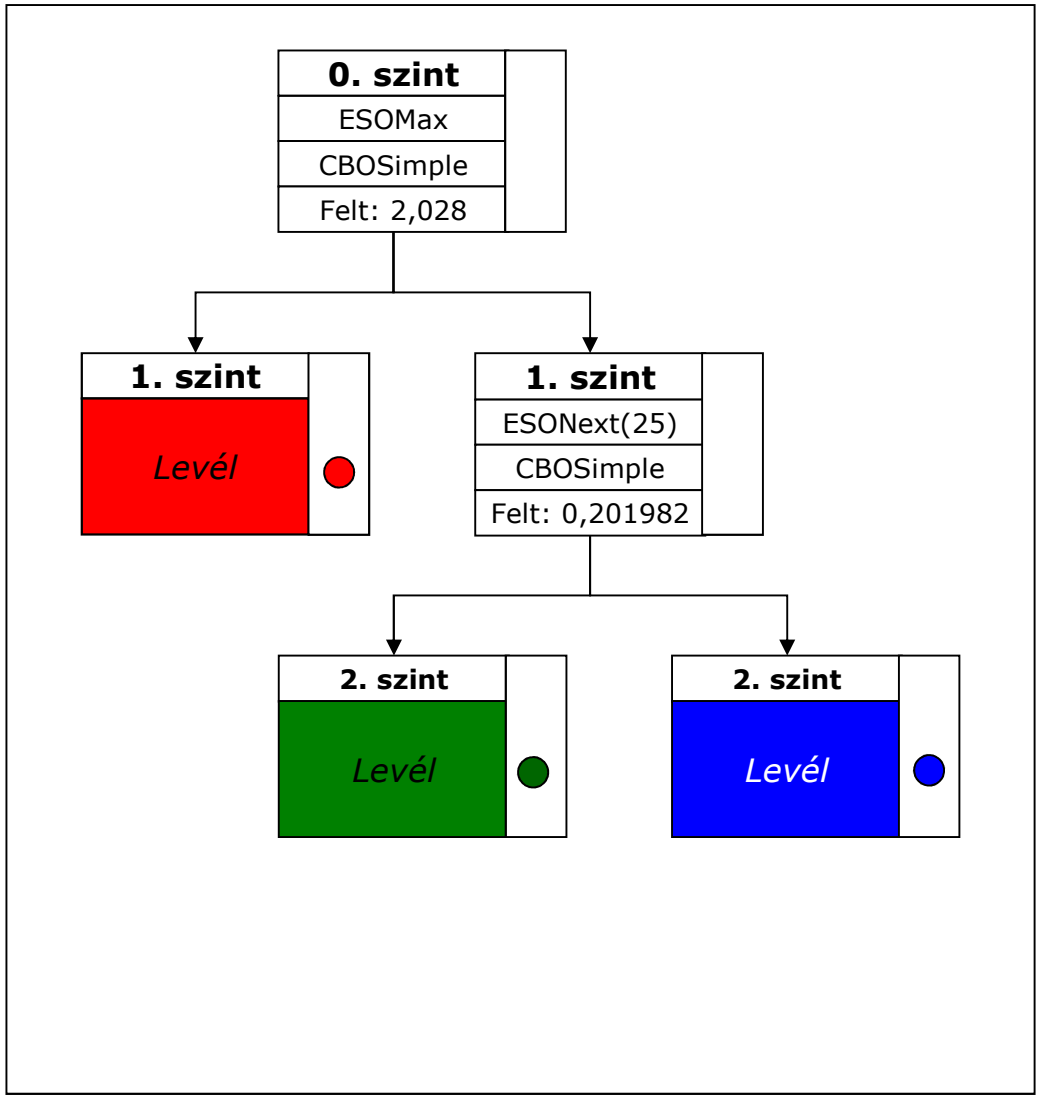


Érték:
1,32432

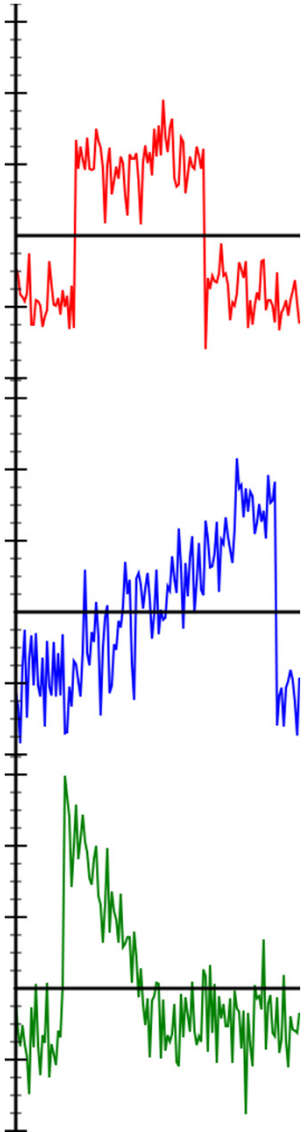


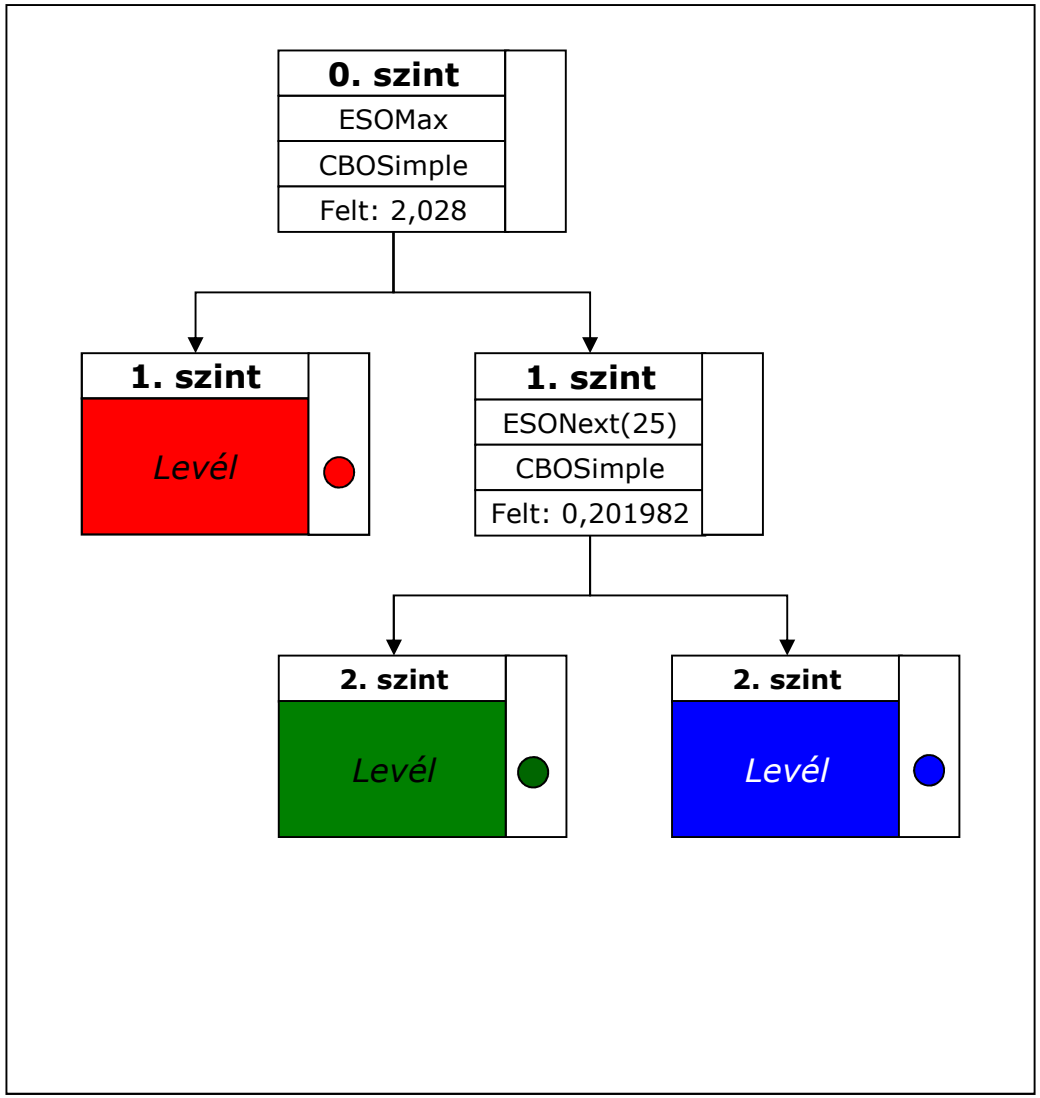
ShiftTree – Osztályozás példa





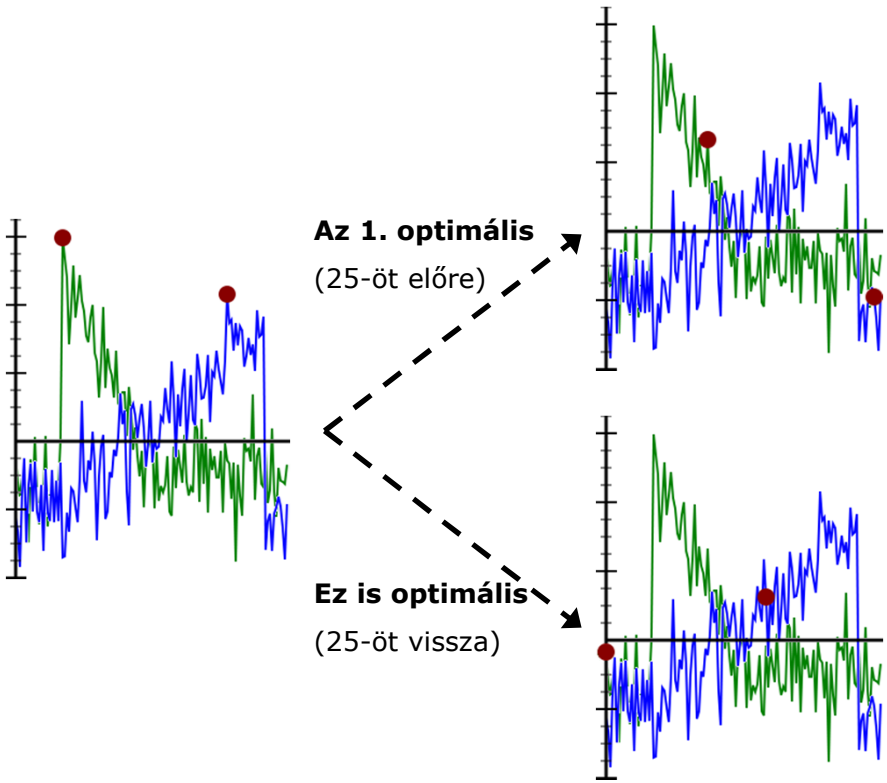
ShiftTree – Osztályozás példa

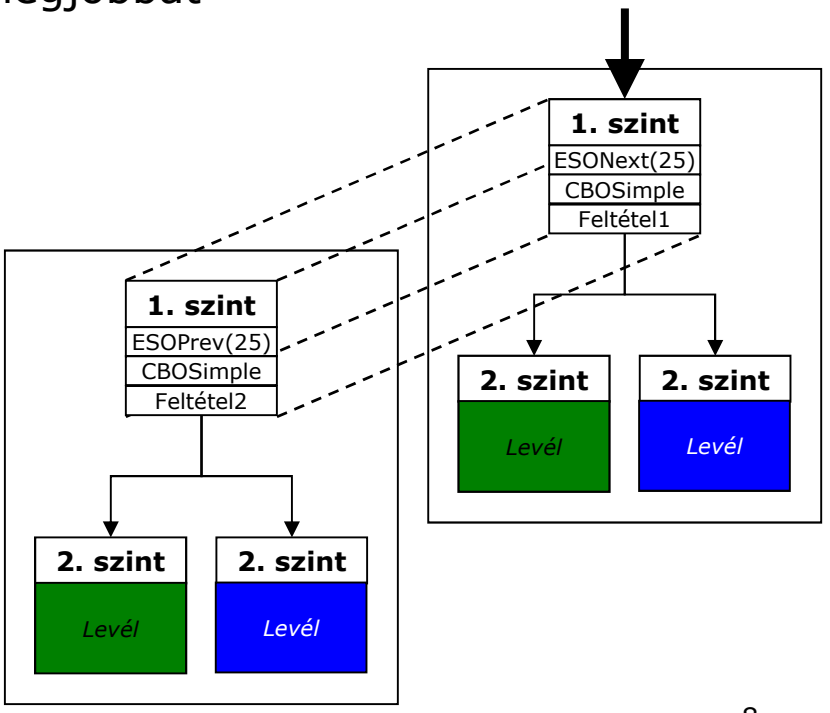




Optimalizálás: többszörös modellezés

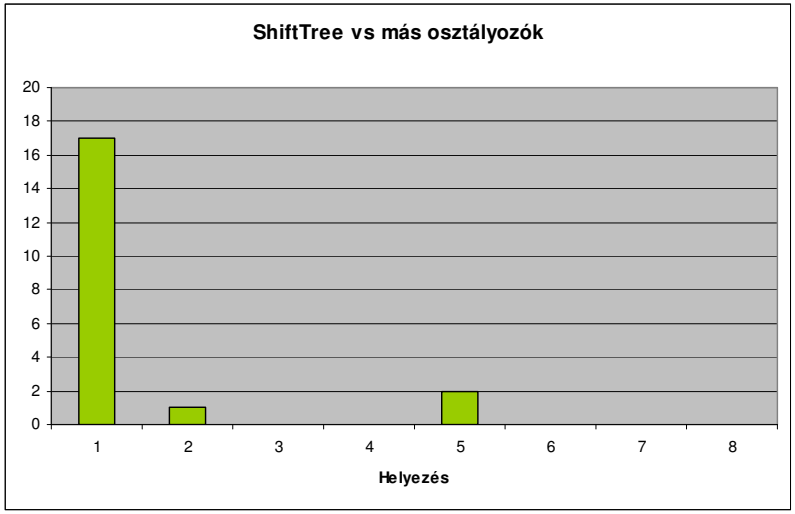
- Több optimális attribútum esetén
  - Az összeset kiválasztjuk
  - Az összes szerint vágunk
  - Többszörös fát építünk
    - De csak ott sokszorozunk, ahol kell, nem az egész fát
  - Egy másik halmazzal kiválasztjuk a legjobbat

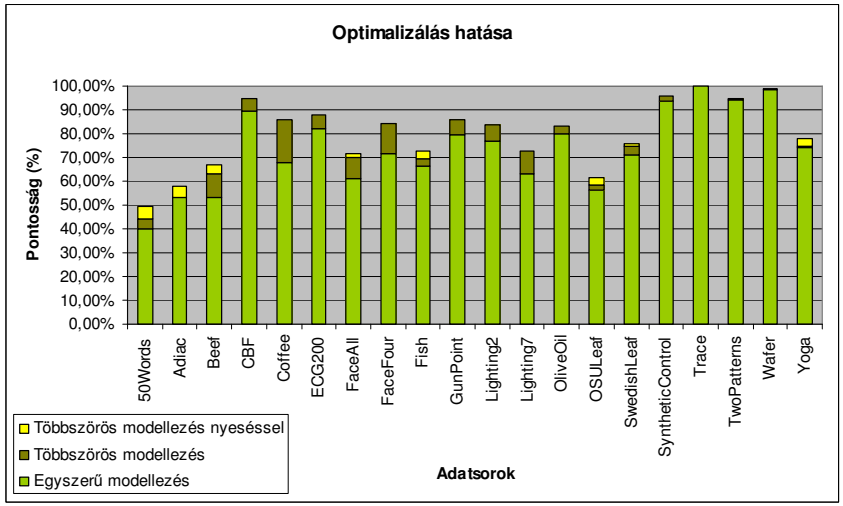




Eredmények: benchmark adatokon

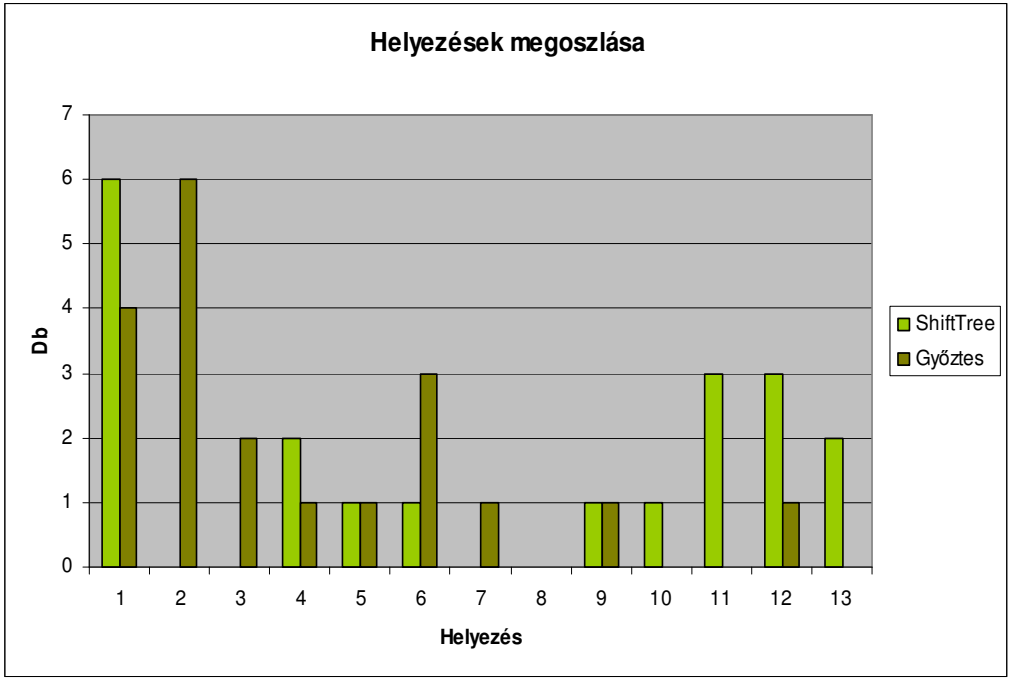
- 20 adatsor különböző területekről
  - Egy változó
  - Eltérő tulajdonságok
  - 7 másik algoritmussal szemben
    - KNN, C4.5 döntési fa, Logistic Model Tree, MLP, SVM, Naív Bayes háló, Random Forest
- Konfiguráció
  - Nincs optimalizálás
  - Legegyszerűbb operátorok
    - Ugrás előre/hátra fix távot
    - Ugrás a következő lokális maximumra/minimumra
    - Ugrás a maximumra/minimumra
    - Pontbeli érték, normális súlyozás, exponenciális súlyozás





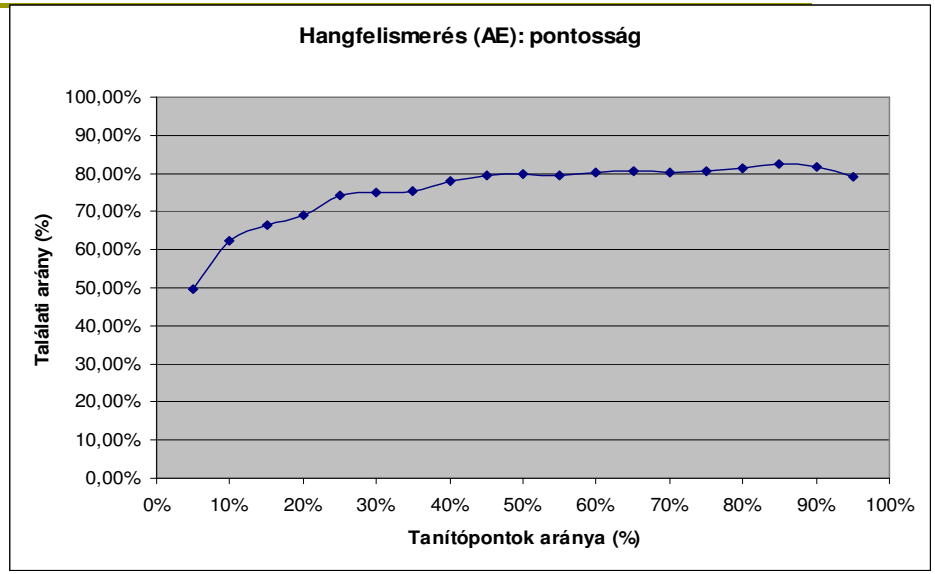
Eredmények: verseny körülmények

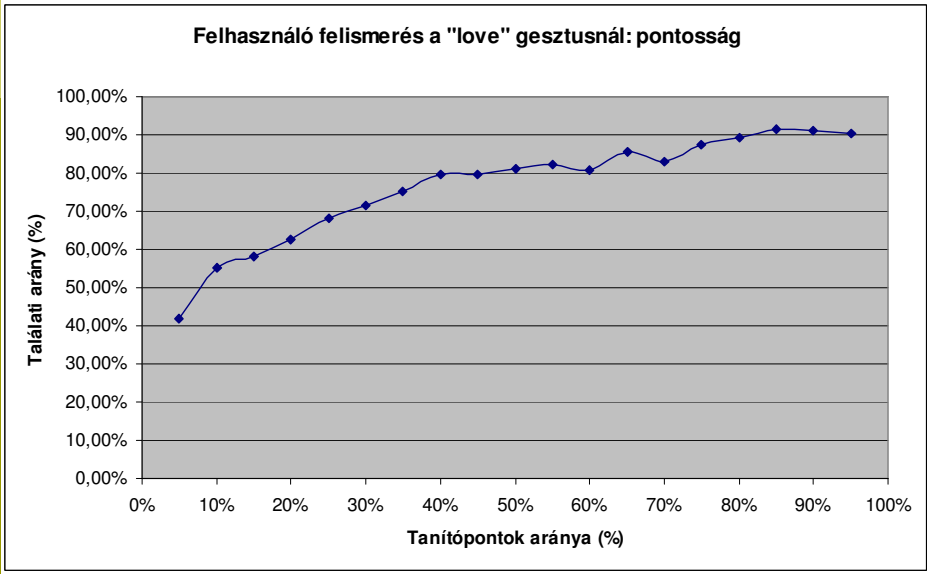
- SIGKDD'07 Time Series Challange adatsorain
  - 20 adatsor
  - Kombinált osztályozók ellen
- Erősebb konfiguráció
  - Fejlettebb operátorok
  - Több futtatás, többségi szavazás
  - De a paraméterek nincsenek finomhangolva
- Eredmények
  - 6 első helyezés (legtöbb)
  - 4 adatsoron még lehetne nyerni
  - 2 adatsoron lehetne javítani
  - 8 adatsoron kevés a tanítóminta
    - Modell alapú algoritmusok itt elvéreznek
  - Összesítésben: 6-8 hely
    - Holtversenyben (a 13-ból)



Alkalmazás: hang- és, gesztusfelismerés

- Személy felismerése az *ae* magánhangzó kiejtése alapján
  - 12 változó
  - 9 személy (osztály)
- Egyszerű operátorok
- Nincs optimalizálás
- Találati arány kellően magas





- Gesztus adatai gyorsulásmérővel
  - 3 változó (koordináta tengelyek)
  - 10 gesztus, 4 felhasználó
  - Kevés adat
- Lehetséges feladatok:
  - Gesztus felismerése
  - Adott gesztusnál a felhasználó felismerése (nehéz feladat)
    - Bonyolult gesztusnál jobb eredmény
    - Kiemelkedő találati arány

Alkalmazás: „Gondolat” felismerés

- EEG hullámok osztályozása
  - Adatsor:
    - Két osztály: FEL, LE
    - 6 változó
  - Jelenleg ~90% körüli pontosság
    - 2003-as versenyen a top3-ban
- Alkalmazás típusai
  - Offline osztályozás
    - Alkatrészek tesztelésének automatikus kiértékelése
  - Stream adatsorban jelek felismerése
    - Még sok nyitott kérdés
    - Folyamatban lévő kutatás

Összefoglalás

- Új idősor-osztályozó: ShiftTree
  - Automatikus
    - Minden egydimenziós idősorral működik
    - Operátorok definiálása a szakértő feladata
      - Nem automatikus
  - Pontos
    - Már egyszerű operátorokkal, optimalizálás nélkül is kellően pontos
    - Optimalizálással kifejezetten hatékony
      - Ha a tanítóminta nem túl kicsi
  - Magyarázó
    - Könnyen értelmezhető modellek, ellenőrizhető
- Legnagyobb előnye: általános
  - Tématerülettől függetlenül hatékonyan használható

További ShiftTree-vel kapcsolatos kutatási anyagok az oldalamon: http://www.hidasi.eu

Köszönöm a figyelmet!