

REZUMAT

**TEHNICI DE DATA MINING
ÎN DIAGNOSTICUL
AUTOMAT**

Docorand: Mircea Sebastian Șerbănescu

Conducător: Prof. univ. dr. Tudor Bălănescu

Domeniu de doctorat: **Informatică**

În epoca medicinei bazate pe dovezi, diagnosticul automat este o alternativă obiectivă la diagnosticul clasic. Printre multe avantaje, din care enumerăm: costuri reduse, timp diagnostic scurt, obiectivitate, ș.a. probabil cea mai puternică trăsătură a diagnosticului automat este aceea că transformă diagnosticul din unul calitativ în unul cantitativ.

Formal, diagnosticul automat, urmează gândirea medicală clinică trecând prin mai multe etape, fiecare fiind pasibilă de perfecționare: pre-procesare, segmentare, extragere de caracteristici, regresie, clusterizare, detecție și clasificare, realizând astfel un proces laborios de distilare a informației numit data mining medical.

Pornind de la probleme medicale reale, cu care am luat contact prin prisma mersului de medic, am propus și testat 13 algoritmi care rezolvă cel puțin una din problemele de data mining amintite.

În ceea ce privește **pre-procesarea** am dezvoltat o nouă tehnică combinată hardware-software pentru normalizarea digitală a colorațiilor lamelor histologice superioară tehnicilor de normalizare software existente atât cantitativ cât și perceptual.

Pentru **extragerea de caracteristici** am dezvoltat 4 algoritmi fiecare aplicându-se unei alte baze de date și având tehnici matematice diferite.

În primul dintre algoritmi am implementat un algoritm de tip box-counting pentru estimarea dimensiunii fractale cu scopul de a ușura clasificarea de imagini de țesut hepatic în țesut normal, cancer primitiv și metastază.

În al doilea algoritm, numit R-VA, am propus o nouă caracteristică pentru imaginile în tonuri de gri. Algoritmul a fost testat atât pe o bază de date publică cât și pe una medicală proprie, iar rezultatele sale au fost superioare algoritmului standard de tip box-counting cu care a fost comparat.

În al treilea algoritm am folosit o parte din caracteristicile extrase din matricea de co-varianță pentru a realiza clasificarea mitozelor din imagini histopatologice de sân.

În al patrulea algoritm am propus o tehnică bazată pe Wavelet-uri pentru detectarea dinamică și neinvazivă a componentelor reflexului de întindere musculară - m1 și m2.

Pentru **segmentare** am dezvoltat un algoritm nesupervizat pentru segmentarea colagenului pe colorația specială tricromică Masson folosindu-mă de spațiul de culoare $L\alpha\beta$ și de un algoritm de tip k-NN. Algoritmul a avut rezultate bune și a scos la iveală o diferență perceptivă între anatomopatologi în ceea ce privește evaluarea colagenului.

Pentru **regresie** am dezvoltat un model de selecție a caracteristicilor bazat pe regresie pentru reducerea numărului de caracteristici și l-am testat cu două rețele neuronale standard. Tehnica asigură un timp de calcul semnificativ mai bun, păstrând totodată același nivel de performanță de clasificare și a fost testată pe o bază de date publică ce urmărește discriminarea tumorilor maligne de sân de cele benigne, precum și detectarea recurenței cancerului.

Pentru **detecție** am dezvoltat un algoritm capabil să evidențieze într-o imagine digitală histopatologică mitozele cu o sensibilitate de 100% și cu o specificitate de 10%. În ciuda valorii scăzute a specificității, anatomopatologii care au evaluat rezultatele au convenit că metoda a fost utilă și că evaluarea rezultatului folosind-o este mai puțin solicitantă decât abordarea clasică.

Pentru **clasificare** am dezvoltat 5 algoritmi fiecare aplicându-se unei alte baze de date și având abordări diferite.

În primul algoritm am propus o rețea neuronală liniară, care a fost comparată cu două rețele standard, folosind două baze de date publice cu privire la detecția și recurența cancerului de sân. Precizia diagnosticului și viteza de calcul ale rețelei liniare au depășit cele două rețele neuronale standard pe ambele baze de date, arătând eficacitatea algoritmului propus.

În cel de-al doilea algoritm am dezvoltat un nou model care are o componentă de algoritm genetic pentru găsirea ponderilor dintre stratul de intrare și cel ascuns și împrumută antrenarea analitică dintr-un extreme learning machine. Modelul a fost testat pe baze de date publice cu cancer de sân și comparat cu trei rețele neuronale clasice, reușind performanțe asemănătoare sau superioare.

În cel de-al treilea algoritm am dezvoltat o versiune nouă a perceptronului multistrat și un mecanism de selecție a caracteristicilor pentru creșterea performanței de clasificare a cancerului de sân. Implementarea vizează utilizarea unui algoritm de selecție a caracteristicilor pentru a elimina redundanța, algoritmi genetici pentru îmbunătățirea procesului de antrenament și conectivitate parțială pentru creșterea vitezei de calcul a rețelei și obține astfel o precizie mai mare a predicției.

În cel de-al patrulea algoritm am propus o tehnică nouă bazându-mă pe metafora evolutivă pentru a obține un model de decizie inteligent cu scop medical. Aplicată unei baze de date de stadializare a fibrozei hepatice, algoritmul a prezentat un comportament stabil și eficient. Tehnica este ușor de adaptat la diferite probleme de decizie, datorită adaptabilității componentelor sale „inteligente”.

În cel de-al cincilea algoritm am propus un nou clasificator de cromozomi bazat pe aprecierea formei din imagini cu cromozomi umani folosind un model k-NN și o bază de date publică. Cea mai bună rată de clasificare a fost obținută pentru doar ($k =$) 3 vecini. Dezavantajul algoritmului este dat de timpul computațional care este de aproximativ 12 secunde pentru un cromozom. Diferite lucrări prezintă metode de clasificare complexe, dar niciuna nu se referă strict la clasificarea formei cromozomilor, aceasta fiind originalitatea tehnicii propuse.

Luând în considerare rezultatele de mai sus și având acces (prin specialitatea mea medicală) la imagini medicale de anatomie patologică, această teză este considerată doar un punct de plecare pentru dezvoltarea ulterioară a abordărilor rapide, precise și obiective, cel mai probabil în tehnici de data mining pentru prelucrarea imaginilor.