

PhD értekezés tézisei

DÖNTÉSI BIZONYTALANSÁG KEZELÉSE DÖNTÉSI FÁKKAL ORVOSI DÖNTÉSTÁMOGATÓ RENDSZEREKBE

Tóth Norbert

Témavezető: Dr. Pataki Béla

Budapesti Műszaki és Gazdaságtudományi Egyetem

Méréstechnika és Információs Rendszerek Tanszék

© 2008 Tóth Norbert

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék
1117 Budapest, XI. Magyar Tudósok körútja 2., I. épület
Phone: (+36 20) 544 5287, Fax: (+36 1) 463 4112,
Email: ntoth@mit.bme.hu, norberttothhun@gmail.com

Tartalomjegyzék

Tartalomjegyzék	3
1 Bevezetés.....	4
2 Bizonytalanság becslése döntési fákkal.....	6
3 Osztályozók kombinálása osztályozási bizonytalanság alapján	12
4 Új tudományos eredmények összefoglalása	15
5 Az eredmények alkalmazása egy valós példán, mammográfia	16
6 Publikációk listája	18
6.1 Nemzetközi folyóiratcikkek.....	18
6.2 Magyar folyóiratcikkek angol nyelven	18
6.3 Magyar folyóiratcikkek magyar nyelven	18
6.4 Nemzetközi konferenciák	18
6.5 Technical Reportok.....	19
6.6 Hazai konferenciák.....	19
7 Irodalomjegyzék.....	20

1 Bevezetés

A nők daganatos megbetegedései között az egyik leggyakoribb az emlőrák. Statisztika szerint minden nyolcadik nőben élete során kifejlődik ez a nemkívánatos betegség. Mivel az emlőrák oka mindmáig ismeretlen, a korai felismerés döntő fontosságú. Korai felismerés esetén az öt éves túlélés esélye 95% körüli.

A mammográfia az egyik legmegbízhatóbb módszer az emlőrák detektálására. Egy mammogram a mellről készült röntgenfelvétel, aminek célja, hogy felfedje a tumorok és ciszták jelenlétét és segítsen eldönteni, hogy azok jó- vagy rosszindulatúak. A kezelés sikere a korai felismerésen múlik és a mammográfia minden más módszernél korábban képes jelezni az eltéréseket.

A mammográfiás szűrésen készített felvételeket jelenleg kizárólag orvosok diagnosztizálják, de ez számítógépes segítséggel is történhetne. Egy teljeskörű mammográfiás szűrés (ami minden nőt érintene 40 év fölött) óriási mennyiségű felvételt jelent. Minden egyes felvétel értékelése nagyon sokáig tart és a folyamat hossza és monotonitása miatt hibákhoz vezethet. Egy támogató rendszer, amely átnézné a felvételeket – kiszűrné a biztos negatívakat és felhívna a figyelmet a gyanúsakra – nagyon hasznos lenne. Ez rengeteg időt takarítana meg és segítene elkerülni a téves diagnózisokat.

Egy a mammogramok szűrésére szolgáló Orvosi Döntéstámogató Rendszer [2] jelenleg is fejlesztés alatt áll a Budapesti Műszaki és Gazdaságtudományi Egyetem Méréstechnika és Információs Rendszerek Tanszékén. A rendszerben több eltérő megközelítésű algoritmus dolgozik egymással párhuzamban, különféle elváltozások után kutatva. Az egyik algoritmus döntési fát használ a szövetminták osztályozására, az emlő textúrájából számított jellemzők alapján. A döntési fák a legelterjedtebb osztályozó modellek közé tartoznak. Kedvező tulajdonságokkal rendelkeznek, mint pl. jó pontosság, értelmezhetőség és magas hatékonyság nagy dimenziószám és adatméret esetén. A disszertáció az eredeti döntési fa elmélet alapjaira épít. Mivel a feldolgozandó mammogramok ~12M pixel felbontású képek, a sebesség fontos tényező. Relatív lassú algoritmusok nem preferáltak, mert a rendszer tanítása, a felvételek kiértékelése és a rendszer kimerítő tesztelése megengedhetetlenül sok időt vehet igénybe. A döntési fák a leggyorsabb osztályozó rendszerek közé tartoznak. Egyszerű felépítéssel rendelkeznek, valamint a tanítás és osztályozás is hatékonyan vihető végbe nagy bemeneti dimenziók esetén is. Másik kedvező tulajdonsága a döntési fáknak a tisztán leírt döntési határfelület, ami az ideális alappá teszi a dolgozatban ismertetett további munkára. A disszertáció az eredeti döntési fa elmélet alapjaira épít.

A kutatási célok és eredmények a disszertációban az Orvosi Döntéstámogató Rendszer fejlesztése során felmerült problémákból és azok megoldásaiból erednek.

Kutatási célok

Több alkalmazás – mint például bizonyos orvosi döntéstámogató rendszerek – megköveteli, hogy az osztályozott minták mellé az algoritmusok egy bizonyossági mértéket is szolgáltatassanak, azt jelezve, hogy mennyire biztosak a diagnózisban. A döntési fák sok kedvező tulajdonságuk ellenére, nem nyújtanak ilyen jellegű információt, nem úgy, mint más osztályozó modellek, mint pl. neurális hálózatok vagy denzitás becslők.

- Az első kutatási cél az eredeti döntési fa elmélet kibővítése a döntési bizonytalanság kezelésére.

Osztályozási problémánál a végső cél egy olyan osztályozó létrehozása, ami a lehető legpontosabban működik. A klasszikus megközelítés szerint több eltérő típusú osztályozót (vagy azonos típusút eltérő paraméterekkel) tanítottak, majd kiválasztották

azt, amelyik a legjobban teljesített. Az osztályozók kiértékelése során kiderült, hogy noha eltérő pontossággal működnek, a rosszul osztályozott minták halmaza nem feltétlenül azonos, de talán még átfedés sincs a köztük. Tehát a különféle osztályozók különféle információkat tanulnak meg az osztályozandó adatokról, és ezt az információt valahogy egyesítve egy olyan osztályozó rendszert lehetne létrehozni, ami teljesítményben felülmúlná a legjobban teljesítő egyedi osztályozót is.

- A második kutatási cél egy új, hatékonyabb osztályozó kombináló rendszer létrehozása. A kutatási cél része az osztályozási bizonytalanságra vonatkozó információ felhasználhatóságának vizsgálata az osztályozók kombinálása során.

A kutatás módszertana

A disszertációban használt vizsgálati módszerek a gépi tanulás, a valószínűségszámítás és statisztika, lineáris algebra és az optimalizáció területéről származnak. Számos módszer átvételre került az irodalomból, ezekre mind hivatkoztam.

A kutatás célja a döntési fák elméletének kibővítése, így a dolgozat erősen épít a döntési fák elméleti alapjaira.

Az általános elméleti eredmények a dolgozatban több szintetikus példán és egy valós alkalmazásban – mammográfiás képfeldolgozás során – is kiértékelésre kerülnek.

A disszertáció felépítése

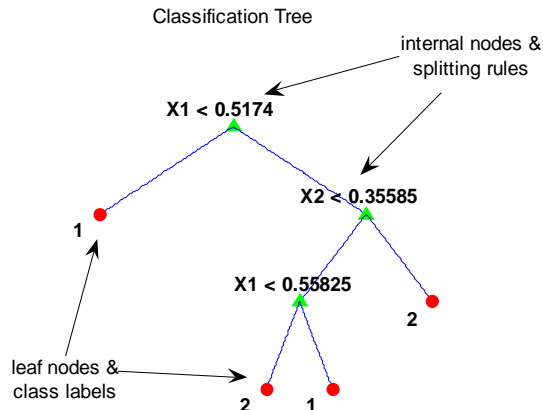
Az elméleti alapok és az irodalomban alkalmazott megközelítések áttekintése után a disszertáció első nagy része (Bizonytalanság becslése döntési fákkal) egy új megközelítést mutat be a döntési bizonytalanság kezelésére döntési fák esetében. A módszer a releváns döntési határfelülettől vett távolság számításán, denzitás becslésen, osztályozási valószínűség és konfidencia becslésen alapul. A módszer nem korlátozódik tengelymerőleges fákra, felhasználható ferde vágású fák esetén is (ahol a döntési hipersíkok nem merőlegesek a tengelyekre), valamint a módszer nem korlátozódik az Euklideszi távolság számítására.

A disszertáció következő nagyobb részében (Osztályozók kombinálása osztályozási bizonytalanság alapján) egy olyan osztályozó kombinálási módszer kerül bemutatásra, ami ezen egyedi osztályozási konfidencia értékeket – amit a kibővített döntési fák számítanak – felhasználva adja meg a végső osztályozást. A javasolt kombinálási forma – osztályozási konfidencia súlyozott többségi szavazás – kedvező tulajdonságokkal rendelkezik a korábbi megközelítésekhez képest. Nincs szükség külső kapuzó hálózatra, a súlyokat adaptívan biztosítják az egyes döntési fák az osztályozó együttesben, valamint új osztályozók dinamikusan adhatók a rendszerhez, anélkül, hogy bármi újratanítást vagy módosítást kellene végrehajtani a meglévő rendszeren.

Az disszertáció utolsó nagyobb része (Az eredmények alkalmazása egy valós példán, mammográfia) elemzi és kiértékeli az ismertetett megközelítéseket egy konkrét valós alkalmazásban: mammográfiás képfeldolgozás során. Értékelésre kerül az egyes (javasolt módszerekkel kibővített) döntési fák kimenete, szintúgy mint az osztályozó rendszer együttes (a javasolt módszerrel számított) kimenete. Több mint 1400 kép került feldolgozásra a megfelelő mélységű kiértékelés érdekében.

2 Bizonytalanság becslése döntési fákkal

A döntési fa típusú osztályozók felügyelt tanulást használva, rekurzívan osztályozva csoportokba osztják a megfigyeléseket, úgy, hogy ezek a csoportok a lehető legnagyobb mértékben különbözzenek egymástól, és maguk a csoportok pedig a lehető leghomogénebbek legyenek.



2.1 ábra. Egy minta döntési fa 2 osztállyal.

A döntési fáknak (amik hipersíkokat használnak a döntések során) két alapvető csoportja létezik: tengelymerőleges fák és ferde vágású fák. A tengelymerőleges fák osztályozó hipersíkjai merőlegesek a bemeneti tér tengelyeire, míg a ferde vágású fák hipersíkjai nem.

Az osztályozási problémák során a végső cél egy olyan osztályozó létrehozása, ami a lehető legpontosabban osztályozza a bemenő mintákat. Mindamelllett sok alkalmazás során szükség van az osztályozás értékelésére, egy bizonyossági érték számítására, hogy az adott osztályozás mennyire megbízható. Ezt a mérést elvégezheti egy az osztályozótól független külső komponens vagy maga az osztályozó is.

A disszertáció második fejezetében egy új módszer kerül ismertetésre, ami kibővíti a döntési fák elméletét lehetővé teszi számukra a döntési bizonytalanság számítását. A javasolt megközelítést alkalmazva a döntési fák képesek egy minden mintára egyedi döntési bizonytalanság érték számítására az osztályozáson felül. A módszer a döntési határfelülettől való távolság számításán és az osztályozási valószínűség becslésén alapul.

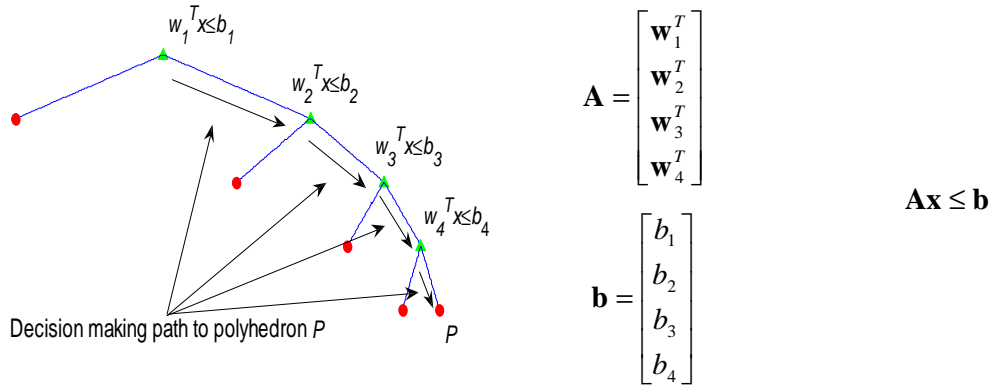
Euklideszi távolság a döntési határfelülettől

A döntési határfelülettől való távolság egy kvadratikus programmal számítható [3], ahol a korlátokat a döntési fa hipersíkjai határozzák meg. A döntési fa minden leveléhez tartozik egy osztály címke, amit a fa minden olyan bemenő mintához hozzárendel, ami ahhoz a levélhez tartozó bemenő térrészben van. A cél a távolság mérése a legközelebbi levélhez, ami más osztállyal rendelkezik, mint a térrész, ahol a bemenő minta található. Ez a probléma egy csoport (minden más osztályú levélhez egy) kvadratikus program megoldásával oldható meg. Egy bemenő pont (s) (Euklideszi) távolsága egy másik levéltől (ami gyakorlatilag egy poliéder, amit egy halom egyenlőtlenség határoz meg) $P = \{x \mid Ax \leq b\}$ R^n -ben $dist(P, s) = \inf\{\|x - s\|_2 \mid x \in P\}$. s P -től való távolságának megtalálásához a következő kvadratikus programot kell megoldani:

$$\text{minimalizálni } f_0 = \frac{1}{2}[\mathbf{x} - \mathbf{s}]^T \mathbf{E}[\mathbf{x} - \mathbf{s}] = \|\mathbf{x} - \mathbf{s}\|_2^2 \quad (2.1)$$

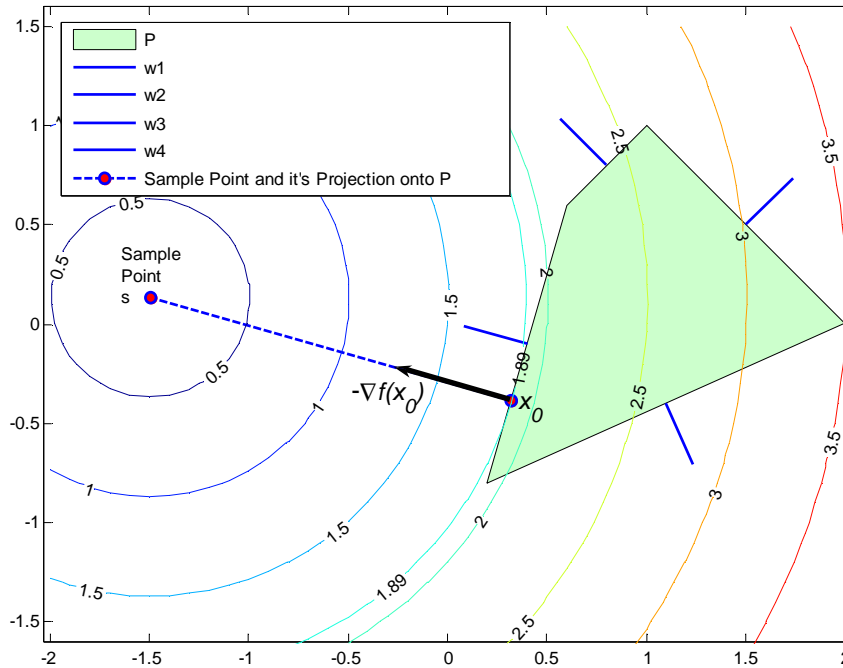
$$\text{feltétellel } \mathbf{Ax} \leq \mathbf{b},$$

ahol \mathbf{s} konstans (a bemenő mintapont koordinátái), \mathbf{E} egy megfelelő dimenziójú egységmátrix (lásd 2.3 ábra). Az \mathbf{A} mátrix és a \mathbf{b} vektor a döntési fa csomópontjaiban található kényszerekből áll a gyökér csomóponttól a kérdéses levélig:



2.2 ábra. A döntési útvonal a gyökér csomóponttól a P levélig. A optimalizációs probléma kényszerei a döntési útvonalon található csomópontokból kerülnek összegyűjtésre.

A döntési fán minden levél egy poliédert határoz meg, ami lefedi a bemeneti tér egy adott részét. Az előbbieknak megfelelően, a döntési határfelülettől való távolság számításához minden más osztállyal rendelkező levélhez ki kell számítani a távolságot. Távolságok azonos osztályú levélhez nem számítanak, mivel azok a régiók azonos osztályhoz tartoznak a bemeneti térben, így nem fenyegetnek a hibás osztályozás lehetőségével. Két azonos osztállyal rendelkező levél lehet szomszéd a bemeneti térben, de nincs valós határfelület köztük.



2.3 ábra. Az S bemenő pont és a P poliéder távolságát – ami megegyezik az S bemeneti pont és az X_0 minimális távolságú vetület pontjának távolságával – az optimalizációs probléma megoldásaként kapjuk.

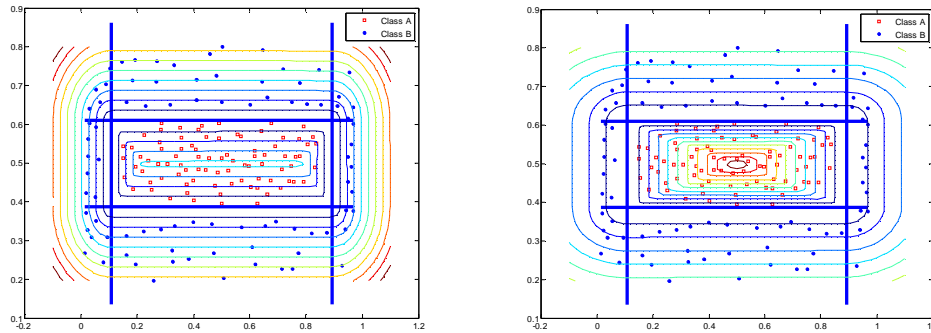
A javasolt megoldás a távolság számítására nem korlátozódik tengelymerőleges fákra, tetszőleges osztályozóval használható, ami hipersíkokat használ elválasztó határfelületeknek.

Mahalanobis távolság a döntési határfelülettől

Az előzőekben ismertetett megoldás – a döntési határfelülettől való távolság számítására – lehetőséget nyújt az Euklideszitől eltérő távolság mértékek használatára, mint pl. a Mahalanobis távolság mérték. Ezek a mértékek nem korlátozódnak azokra, ahol a minimális távolságú vetület azonos. Ez azt jelenti, hogy egy adott s bemenő mintához egy poliéderben (döntési fa levélben) a legközelebb eső pont nem feltétlenül azonos eltérő távolság mértékek esetén. A Mahalanobis távolságot P. C. Mahalanobis [4] vezette be. Ez egy szórás mértékű távolság, ahol a bemenő változók közötti korrelációk is figyelembe vannak véve. A Mahalanobis távolság két véletlen x és y vektor között Σ kovarianciamátrixszal:

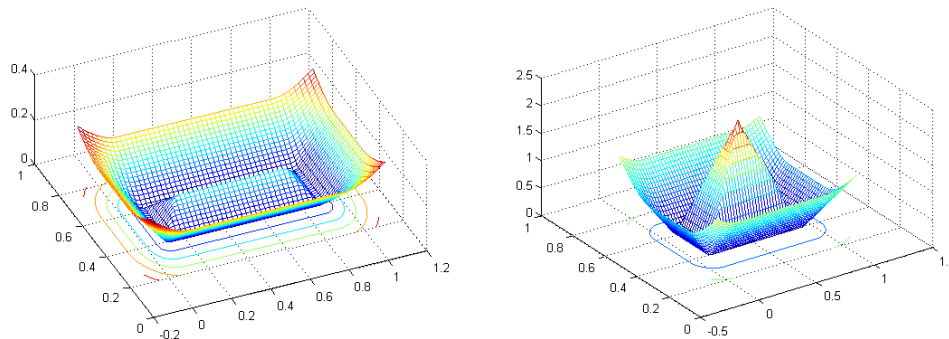
$$dist(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})} . \quad (2.2)$$

A Mahalanobis mértéket használva a távolság definíciójára a bemenő minta távolságának számítására a döntési határfelülettől azon a hipotézisen alapul, hogy nagyobb szórásirányokba nagyobb távolság szükséges a határfelülettől azonos osztályozási bizonyossághoz, mint kisebb szórásirányokba. Ezt illusztrálják a következő ábrák. Az adatpontok a középső négyzet alakú térrészben nagyobb szórással rendelkeznek vízszintes irányban, mint függőleges irányban. Ez arra utal, hogy vízszintes irányban messzebb kell lennie egy osztályozott pontnak a döntési határfelülettől, hogy azonos osztályozási bizonyosságot kapjon.



2.4 ábra. Két osztályos osztályozási probléma és egy döntési fa négy vágással (vastag vonalak). A szaggatott kontúrok a döntési határfelülettől való távolságot illusztrálják. Euklideszi távolság balra, Mahalanobis jobbra.

A Mahalanobis távolság szerint nagyobb szórási irányokban (ebben az esetben a vízszintes tengely) a távolság lassabban emelkedik, mint a kisebb szórási irányokban (ebben az esetben a függőleges tengely). A 2.4 ábrán a távolság kontúrok láthatók Mahalanobis és Euklideszi távolság esetében. A következő 2.5 ábrán a döntési felülettől vett távolságértékek egy hálóként láthatók. Látható, hogy Mahalanobis esetben a távolság gradiense függőleges irányban sokkal gyorsabban emelkedik, mint vízszintes esetben (kisebb szórási irány), míg Euklideszi esetben minden irányban azonos gradienssel nő a távolság.



2.5 ábra. Távolság értékek a döntési határfelülettől. Euklideszi balra, Mahalanobis jobbra.

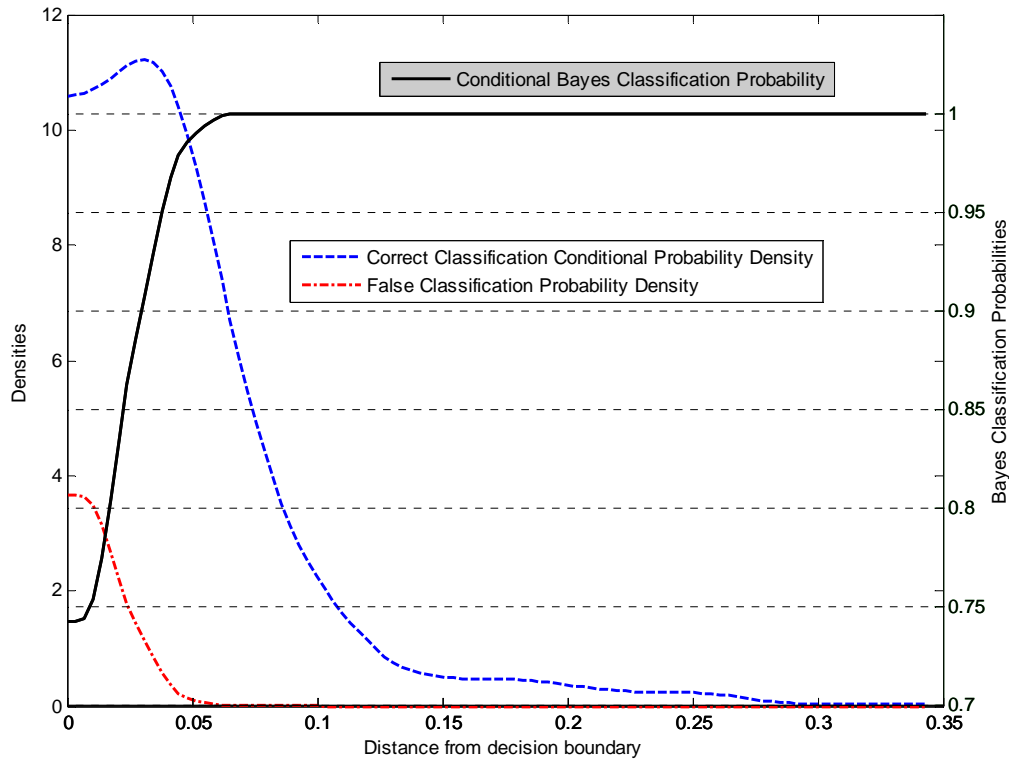
Feltételes osztályozási valószínűségek

A távolságok számítása után a javasolt megoldás denzitás becslést [5] alkalmaz a helyes / hibás osztályozási (vagy osztály) valószínűségek meghatározására. A módszer a távolság információ felhasználásával kiküszöböli a (legtöbb) denzitás becslő legnagyobb problémáját, az exponenciális robbanás veszélyét.

A távolság számítás segítségével a módszer a bemenő mintapontokat egy 1 dimenziós távolság térbe képi le, ahol az egyetlen információ a döntési határfelülettől vett távolság. Így a denzitások becslését már csak ezen az 1 dimenziós adathalmazon végezve a módszer elkerüli az exponenciális számításidőt és tárigényt.

A módszer megoldást ad a döntési határfelületek közelében a 0 távolságok kezelésére is. A megoldás a tükrözési technika segítségével kisítmítja a denzitásgörbét a határfelületek közelében, kiküszöbölve a meredek csúcsokat és völgyeket, biztosítva $\hat{f}'(0+) = 0$. Ez a kedvező tulajdonság robusztus denzitás becsléseket és osztályozási valószínűségeket eredményez. A javasolt megoldás alkalmazható függetlenül attól, hogy milyen bonyolult levél struktúra vagy hány hipersík határolja a leveleket.

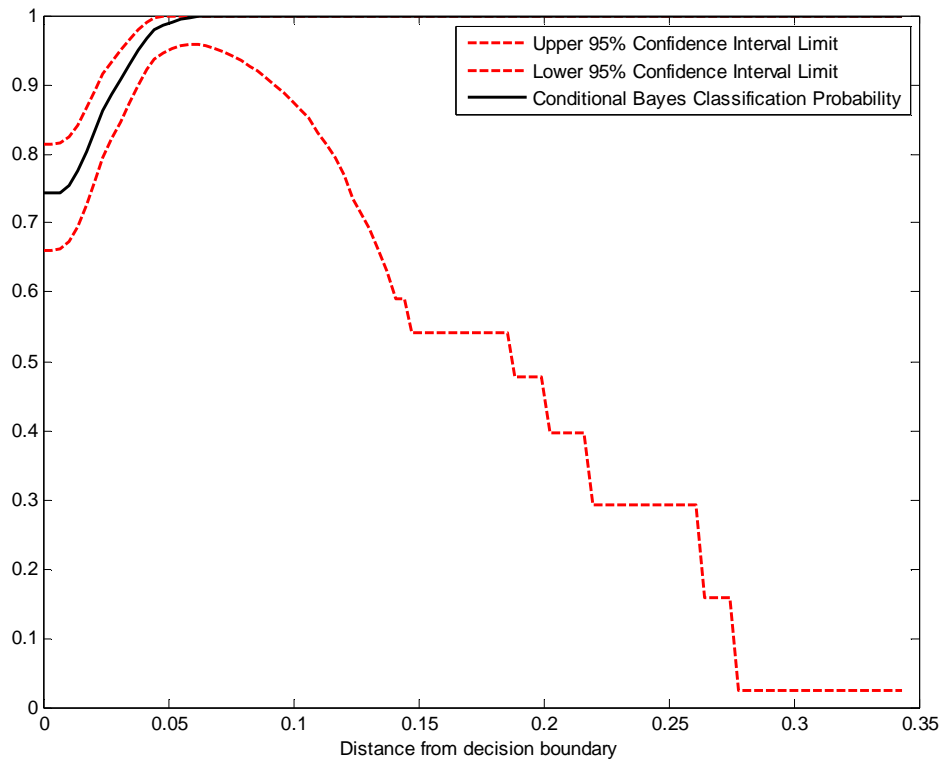
A javasolt megoldás a Bayes szabály [6] segítségével határozza meg a denzitásokból a helyes osztályozási valószínűséget.



2.6 ábra. Helyes és hibás osztályozási valószínűségi denzitások, valamint a helyes osztályozás Bayes valószínűsége.

Konfidencia intervallum

A fejezet további szakaszában egy további javaslat kerül ismertetésre a konfidencia intervallum számítására [7] az előzőekben meghatározott helyes osztályozási valószínűséghez. A megközelítés lényege, hogy minden osztályozó felfogható egy fekete dobozként, ahol a sikeres osztályozások száma a binomiális eloszlást követi (amihez konfidencia intervallum számítható az F eloszlás felhasználásával [8]). A javasolt megoldásban a sikeres osztályozások száma a döntési határfelülettől vett távolság függvénye. Az osztályozási valószínűségekhez tartozó konfidencia intervallum számítására meghatározott távolságoknál szükség van az adott távolságban lévő tanító pontok számára (a konfidencia intervallum a helyes osztályozási valószínűségtől és a helyes osztályozási valószínűség becsléséhez felhasznált mintaszámtól függ). A módszer a rendelkezésre álló minták számát a denzitásokból becsli.



2.7 ábra. 95%-os konfidencia intervallum a helyes osztályozási valószínűsége egy minta döntési fára, minden d távolságra eső mintához a döntési határfelülettől.

A kapott konfidencia intervallum adott távolságra a döntési határfelülettől a helyes osztályozási valószínűségtől és a helyes osztályozási valószínűség becslésére használt mintaszámtól függ.

Értékelés

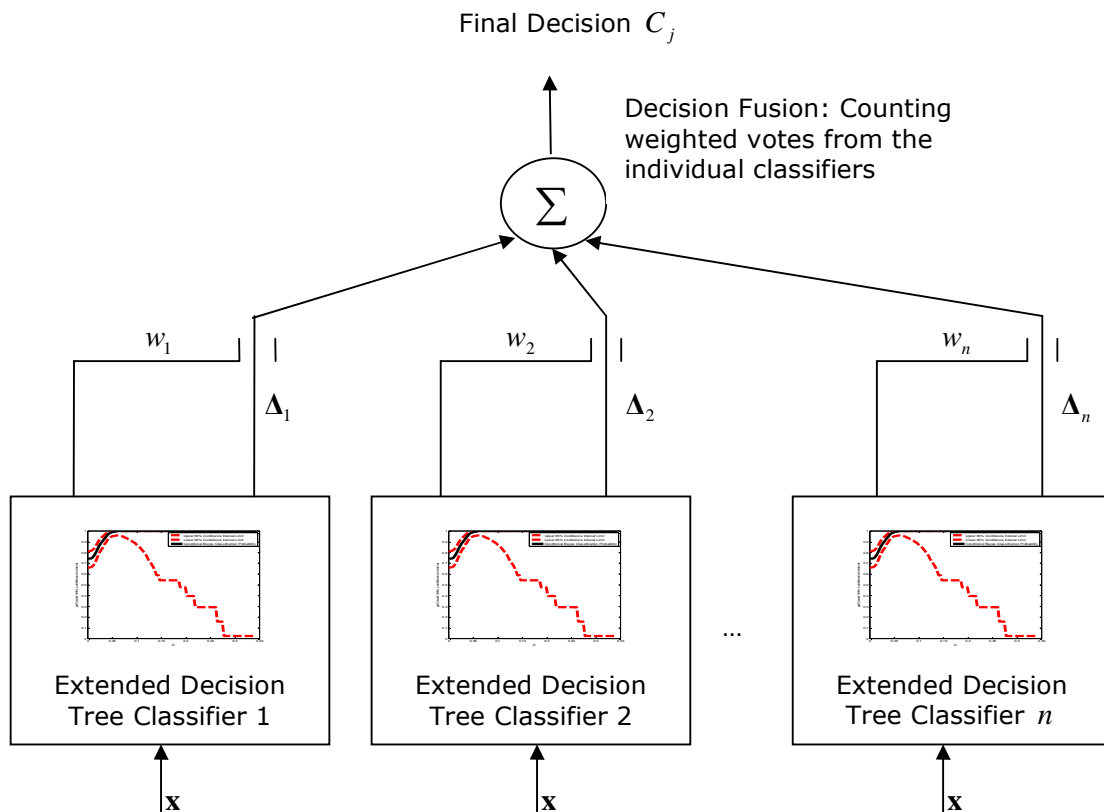
A javasolt megközelítések három minta adathalmazon kerülnek kiértékelésre a fejezetben. Az első alkalmazási példában egy bizonyossági határ kerül meghatározásra, ami alatt a rendszer visszautasítja az osztályozást. Az eredmények szerint a javasolt megoldás jelentősen tudja javítani az osztályozó pontosságát, azon az áron, hogy visszautasítja a bementő pontok egy kockázatos halmazát. A második vizsgálatban az osztályozók ROC analízise történt meg. A javasolt megoldással kibővített döntési fák olyan ROC görbéket produkáltak, amik dominálják a Laplace korrekcióval nyert görbéket, valamint finomabb TP / FP választást tesznek lehetővé, mivel minden mintaponthoz egyedi értéket számolnak, ellentétben a Laplace korrekcióval (ami a levél valószínűségek finomításán alapszik).

3 Osztályozók kombinálása osztályozási bizonytalanság alapján

A disszertáció harmadik fejezetében egy új megközelítés kerül bemutatásra (osztályozási bizonytalansággal súlyozott többségi szavazás) a döntési fa típusú osztályozók kombinálására. A módszer alapja az „összeg” és a többségi szavazás szabály [9], valamint levezetése során felhasználásra kerültek az előző fejezet (Bizonytalanság becslése döntési fákkal) eredményei.

A javasolt osztályozási konfidencia súlyozott többségi szavazás

Az előző fejezet adott egy módszert az osztályozási valószínűséghez tartozó konfidencia intervallum számítására. Ez a konfidencia reprezentálja az osztályozó bizonyosságát az osztályozásban és egyedi minden osztályozott mintára (a minta döntési határfelülettől való távolságának függvénye). Ezen konfidencia értékeket felhasználva minden osztályozó be tudja állítani a saját súlyát a végső döntés meghozatalában. Így a súlyok minden osztályozóra adaptívan változnak a bemeneti minta függvényében, attól függően, hogy az egyes osztályozók mennyire biztosak az osztályozásban. Az olyan osztályozók, amik nagy konfidenciával osztályozzák az aktuális mintát magasra állítják a saját súlyukat, valamint azok az osztályozók, amik alacsony konfidenciával osztályozzák az aktuális mintát kis súlyt állítanak be maguknak. Az – a javasolt módszerrel kibővített döntési fákat tartalmazó – osztályozó együttes architektúrája látható a következő ábrán (3.1 ábra):



3.1 ábra. A javasolt súlyozott szavazó rendszer. A súlyokat a kibővített döntési fák adaptívan állítják be a számított osztályozási konfidenciának megfelelően.

A végső osztályt a következő képlet határozza meg:

legyen $Z \rightarrow C_j$, ha

$$\sum_{i=1}^R \Phi_i(\mathbf{x}) \Delta_{ji} = \max_{k=1}^m \sum_{i=1}^R \Phi_i(\mathbf{x}) \Delta_{ki}, \quad (3.1)$$

ahol

$$\Delta_{ki} = \begin{cases} 1 & \text{Ha az } i\text{-ik osztályozó } k \text{ osztályra szavaz} \\ 0 & \text{egyébként.} \end{cases} \quad (3.2)$$

és $\Phi_i(\mathbf{x})$ az i -ik osztályozó súlya a rendszerben, azaz az i -ik osztályozó konfidencia értéke. Ez egy bizonyossági érték, amit az osztályozó csatol a minta osztályához, azt mutatva, hogy az osztályozás helyességének valószínűsége ennél nem kevesebb (egy megadott konfidencia szinten). A konfidencia szint beállítása a fák tanítása előtt történik. Amennyiben nem szerepel külön, a dolgozatban mindenhol 95%-os konfidencia intervallumok szerepelnek.

A javasolt szavazó rendszer adaptívan súlyozza az osztályozókat az együttesben, attól függően, hogy azok milyen konfidenciával osztályozzák a bemenő mintát. Így egy bemenettől függő, adaptív, önszerveződő osztályozó kombináló rendszert kapunk:

- **Bemenettől függő:** minden osztályozó becsli a saját osztályozási konfidenciáját, ami a bemenő minta döntési határfelülettől mért távolságának függvénye.
- **Adaptív:** mivel a súlyok az osztályozók kimeneteinek kombinálása során az osztályozók konfidenciáiból kerülnek meghatározásra, amik a függnek a bemeneti mintától, így a súlyok is a bemeneti minta függvényei lesznek.
- **Önszerveződő:** a súlyokat az osztályozók osztályozási konfidenciái határozzák meg. Így az architektúrában nincs szükség egy külső kombináló hálózatra vagy logikára. Az architektúra lehetővé teszi, hogy a rendszerhez dinamikusan új osztályozókat adjunk hozzá (vagy vegyünk ki) anélkül, hogy bármilyen kombináló logikát újra kelljen tanítani, vagy módosítani. Ha egy új osztályozó kerül a rendszerbe, semmilyen komponensen nem szükséges változtatást végrehajtani.

A kompetencia területről kiterjedő döntési határfelületek problémája

Az osztályozókat tanító halmazon tanítják, kompetenciájuk a tanító halmazzal lefedett bemeneti térrészre (kompetencia régió) korlátozódik. A disszertációban javasolt megoldás lehetővé teszi az osztályozók számára, hogy egy konfidencia értéket számítsanak minden bemeneti mintára. Ez az érték függ a döntési határfelülettől egy adott távolságra lévő tanítópontok számától.

Ennek az értéknek korrekt, értelmezhető jelentése van az osztályozók kompetencia régióiban. Mindazonáltal a döntési határfelületek kinyúlhatnak az osztályozó kompetencia régiójából, ahol az osztályozási konfidenciák elvesztik jelentésüket. Ez a jelenség abból következik, hogy a konfidencia érték az osztályozó pontosságától és a pontosság becsléséhez használt mintaszámtól függ, amik egy 1 dimenziós leképzésen (a döntési határfelülettől való távolságra) keresztül állnak rendelkezésre. Az olyan döntési határfelületek, amik „kilógnak” a kompetencia régióból, hamis osztályozási bizonyosságokat eredményezhetnek olyan régiókban, ahol egyáltalán nem áll rendelkezésre tanító minta.

Az olyan minták, amik távol esnek a tanító pontok régiójától, de közel vannak a döntési határfelülethez, kis távolság értéket fognak kapni. Mivel az osztályozási konfidenciák függnek a számított távolságoktól, ezek a pontok nagy konfidenciával lesznek osztályozva, noha valójában kilógó pontok. A régiójukban nem volt tanító minta, egy ilyen más kompetencia régióba tartozó döntési fa általi osztályozásban nem szabad megbízni.

A kiterjedő határfelületek problémája megoldható egy korrekcióval a döntési határfelülettől való távolság számítása után. A javasolt megoldásban egy mesterséges döntési felületet képezünk a tanító minta köré. Később, amikor a bemeneti minta távolsága számításra kerül a döntési határfelülettől, a korrekt távolság a maximuma a valós döntési határfelülettől való távolságnak és ennek a mesterséges kompetencia régiót határoló felülettől vett távolságnak:

$$d(\mathbf{x}) = \max\{d(\mathbf{x}, \text{döntési határfelület}); d(\mathbf{x}, \text{kompetencia határ})\}. \quad (3.3)$$

A megoldás kiküszöböli a tanító minták által lefedett régióból kiterjedő döntési határfelületek által okozott problémát. Az olyan minták, amik távol esnek a kompetencia régiótól, de közel egy döntési határfelülethez, nagy távolságértékeket fognak kapni és alacsony osztályozási konfidenciákat.

Értékelés

A javasolt megközelítés összehasonlításra került két elterjedt és elfogadott (különböző elven működő) osztályozó kombináló módszerrel:

- A Mixture of Experts struktúra (MOE) [10], és
- Bagging [11].

Ez a két megoldás eltérő megközelítések szerint épít és kombinál osztályozókat. A MOE struktúra célja a bemeneti tér partícionálása és osztályozók rendelése a partíciókhoz. Amikor bemeneti minta osztályozásra kerül, az a szakértő kapja a legnagyobb súlyt a végső döntés kialakítása során, amelyik a legkompetensebb az adott régióban. A szakértők súlyozását egy külső kapuzó hálózat valósítja meg, a bemeneti mintának megfelelően.

A bagging egy más megközelítést alkalmaz. Ugyanazon tanító halmazból újramintavételezéssel (bootstrap) új tanító halmazokat alakítunk ki és ezeken tanítjuk az osztályozókat. Ebben az esetben minden osztályozó ugyanabban a régióban helyezkedik el. A végső döntés meghozatala többségi szavazással történik.

Az összehasonlítások megmutatták, hogy a javasolt megoldás alkalmazható függetlenül attól, hogy az osztályozók kompetencia régiói átfednek-e a bemeneti térben. A javasolt megoldás imitálni tudja a MOE rendszer működését, mivel a számított konfidencia értékek felfoghatók a kapuzó hálózat kimeneteinek, amik a bemeneti teret partícionálják. A bagginggel történő összehasonlítás során kiderült, hogy a módszer abban az esetben is alkalmazható, ha az osztályozók ugyanazt a régiót foglalják el a bemeneti térben és kimeneteik nagymértékben korrelálnak. A vizsgálatok megmutatták, hogy a javasolt kombináló struktúra hatékonyan alkalmazható anélkül, hogy az osztályozók elhelyezkedésére vagy függetlenségére figyelni kellene.

4 Új tudományos eredmények összefoglalása

Az első téziscsoport egy új megközelítést mutat be a pontosabb osztályozási valószínűségek számítására döntési fa típusú osztályozók esetében. A módszer alapja a távolság számítása a releváns döntési határfelülettől. A második téziscsoport egy új módszert javasol a döntési fa típusú osztályozók kombinálására, felhasználva az első téziscsoport eredményeit.

1. Téziscsoport: Egy új módszert adtam az osztályozási valószínűségek pontosabb számítására, felhasználva a bemeneti minta távolságát a döntési fák által meghatározott releváns döntési határfelülettől. A téziscsoport eredményei a következő publikációkban jelentek meg: [p 1][p 2][p 4][p 8][p 9][p 10][p 12][p 14][p 15].

- 1.1. Úgy fogalmaztam meg a problémát, hogy egy kvadratikus programmal megoldhatóvá vált a bemenő mintának döntési fa releváns döntési határfelületétől való távolságának kiszámítása. A módszer nem korlátozódik a tengelymerőleges fákra, tetszőleges döntési fával (vagy osztályozóval) alkalmazható ahol, a hipersíkok osztják fel a bemeneti teret.
- 1.2. Megmutattam, hogy a döntési határfelülettől való távolság számítására javasolt módszer különféle távolság mértékekkel is használható az Euklideszin kívül, amelyeknél a minimális távolságú vetület nem feltétlenül azonos az eltérő távolság mértékekre. Bemutattam a Mahalanobis távolság alkalmazását.
- 1.3. Az osztályozási valószínűségek meghatározására a bemenő minták döntési határfelülettől való távolságának függvényében történő denzitás becslését javasoltam. Megoldási javaslatot adtam a nulla távolság kezelésére a döntési határfelületek közelében.

2. Téziscsoport: Kifejlesztettem egy megoldást az osztályozási konfidenciák számítására az első téziscsoport által nyújtott távolság feltételes osztályozási valószínűségekhez, valamint megmutattam, hogy ez az osztályozási konfidencia hatékonyan felhasználható osztályozók kombinálása során. A tézis csoport eredményei a következő publikációkban jelentek meg: [p 1][p 2][p 8][p 9][p 10][p 12][p 14][p 16].

- 2.1. Kifejlesztettem egy módszert konfidencia intervallum számítására a távolság feltételes osztályozási valószínűséghez.
- 2.2. Egy adaptív módszert javasoltam döntési fa típusú osztályozók kombinálására, felhasználva az osztályozási konfidencia információt. A létrejött osztályozó rendszer a konfidencia súlyozott többségi szavazás, a hagyományos többségi szavazás kibővítésének tekinthető.
- 2.3. Megoldást adtam az osztályozók kompetenciaterületéről kinyúló döntési határfelületek problémájának kezelésére az osztályozók kombinálása során.

5 Az eredmények alkalmazása egy valós példán, mammográfia

A disszertáció általános eredményei a negyedik fejezetben egy valós alkalmazás során kerülnek kipróbálásra, mammográfiás képfeldolgozás során. Egy minta alkalmazás során bemutatásra kerülnek az osztályozási konfidencia számításával (2. fejezet) kibővített döntési fa osztályozók használatának előnyei a mamogramok elemzése során. A mamogramok részekre bontása után minden rész több döntési fával kerül osztályozásra, majd az eredmények integrációja a javasolt konfidencia súlyozott többségi szavazással történik (3. fejezet).

Az osztályozó rendszer kimenete egy szavazólap, ahol minden képrészlet rendelkezik egy szavazati számmal, ami az adott régió gyanúságát mutatja. Minden döntési fa egy -1 vagy +1 szavazat értéket ad, attól függően, hogy az adott képrészlet háttérszövet vagy kóros szövet (folt). Ezen szavazatok összege (súlyozva az osztályozási bizonytalansággal súlyozott szavazás esetén) alkotja a szavazólapokat.

A rendszer több mint 800 pozitív és több mint 600 negatív képen lett tesztelve. A fejezet bemutat egy kiértékelési eljárást, valamint a javasolt megoldás összehasonlításra kerül a legelterjedtebb kombináló rendszerrel a többségi szavazással (különösen a standard döntési fák esetében, ahol csak osztályozási információk állnak rendelkezésre). Az eredmények kvalitatívan és kvantitatívan is értékelésre kerülnek. A kvalitatív eredményeket alátámasztják a kvantitatív eredmények, amik szignifikáns javulást mutatnak a legelterjedtebb módszerrel, a többségi szavazással szemben.

A fejezetben bemutatott rendszer egyes részei megjelennek a Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszékén fejlesztett Orvosi Döntéstámogató Rendszerben [2]. A fejezetben bemutatott eredmények összehasonlítása nem lenne korrekt komplett mammográfiás képfeldolgozó rendszerekkel a következő okok miatt:

- A fejezet célja a disszertáció általános eredményeinek alkalmazhatóságát bizonyítani és nem pedig a lehető legjobb találati arányt elérni mammogramokon. A fejezetben ismertetett rendszerből hiányoznak olyan elő- és utófeldolgozó részek, amik szignifikánsan befolyásolhatják a rendszer teljesítményét.
- Kereskedelmileg elérhető rendszerek mint pl. a Second Look vagy az R2 titkosak, információ gyakorlatilag nem érhető el a működésük részleteiről. Általánosságban elmondható, hogy még olyan más rendszerekkel való összehasonlítás, mint pl. az Analogikai mammográfiás szűrőállomás [12] - [13], ahol valamilyen mennyiségű információ elérhető, is helytelen anélkül, hogy meglennének pontosan ugyan azok a tesztek képek, valamint a pontos kiértékelési szabályok.

Eredmények pozitív felvételeken

A rendszer 10 pozitív felvételeket tartalmazó DDSM kötetben [14] került kiértékelésre, összesen 829 képen. Minden pozitív felvétel tartalmazott legalább egy elváltozást. Az eredmények szerint a javasolt osztályozási konfidenciával súlyozott szavazást használva a többségi szavazással szemben a döntési fák kombinálásra 6.6%-os TP arány javulás érhető el (kép szinten), amennyiben az algoritmus csak a leggyanúsabb részt jelöli meg a szavazólapon, mint foltjelöltet (5.1). Más szemszögből nézve a javasolt megoldással 68%-os (132 fals pozitív többségi szavazással és 41 a javasolt módszerrel) FP arány javulás érhető el, amennyiben minden egybefüggő régiót megjelöl az algoritmus foltjelöltként egészen addig, amíg a küszöb annyira le nem csökken, hogy a valós folt is belekerüljön a jelöltek közé (5.1 táblázat).

5.1 táblázat. Összesített eredmények 10 DDSM kötetből. 829 felvétel.

	Többségi szavazás	Konfidencia súlyozott szavazás	Javulás
# Fals pozitív	132	42	68,4%
TP találati arány a legmagasabb konfidencia / szavazat értéknél	88,4%	95,0%	6,6%

Eredmények negatív felvételeken

A rendszer 2 negatív felvételeket tartalmazó kötetben kerül kiértékelésre, összesen 611 felvételen. Először egy átlagos vágási határ került meghatározásra a 829 pozitív képen, mind a többségi szavazáshoz, mind a javasolt konfidencia súlyozott szavazáshoz. Ezek az átlagos vágási határok az egyes képeken található azon szavazólap értékekből kerültek meghatározásra, ahol a valódi folt megtalálható volt. Ezen átlagos gyanúsági küszöbökkel (1 a többségi szavazáshoz, 1 a konfidencia súlyozott szavazáshoz) a negatív felvételek szavazólapjai vágásra kerültek. A küszöbözés után az egybefüggő részeket megszámlálva megkapható a fals pozitív jelölések száma képenként. Minél jobban teljesít az algoritmus, annál kisebb ez a szám. Az eredmények szerint a javasolt konfidencia súlyozott kombinálást használva a többségi szavazással szemben 7.5% fals pozitív arány javulás érhető el a negatív tesz felvételeken (5.2 táblázat).

5.2 táblázat. Összesített eredmények 2 DDSM kötetből. 611 felvétel.

	Többségi szavazás	Konfidencia súlyozott szavazás
# Fals pozitív	4437	4103
Átlagos fals pozitív / felvétel	7,26	6,72

Összegzés

A disszertáció ezen fejezetében egy mammográfiás képfeldolgozó rendszer részei – osztályozás és szavazás – kerültek kiértékelésre. Az eredmények megmutatták, hogy a javasolt megközelítéseket használva szignifikáns pontosság javulás érhető el. Mindazonáltal ezen analízis nem hasonlítható komplett mammográfiás képfeldolgozó rendszerekhez, amik további – ezen vizsgálat során nem alkalmazott – komponenseket is használnak. Ilyen komponensek pl. az előfeldolgozás, utófeldolgozás, a felvételek együttes elemzése, stb. Ezen komponensek használata szignifikánsan befolyásolhatja a rendszer végső teljesítményét. Az elemzés célja, hogy bemutassa a disszertáció általános eredményeinek alkalmazhatóságát. Természetes, hogy a bemutatott rendszer tovább javítható az említett rendszerkomponensek használatával, amennyiben a végső cél a minél jobb találati arány elérése mammogramok osztályozása során.

6 Publikációk listája

6.1 Nemzetközi folyóiratcikkek

- [p 1] N. Tóth, B. Pataki: "Classification Confidence Weighted Majority Voting Using Decision Tree Classifiers", *International Journal of Intelligent Computing and Cybernetics*, accepted, to appear in June, 2008.
- [p 2] N. Székely, N. Tóth, B. Pataki: "A Hybrid System for Detecting Masses in Mammographic Images", *IEEE Transactions on Instrumentation and Measurement*, Vol. 55, No. 3, June 2006, pp. 944-952.
- [p 3] L. Lasztovicza, B. Pataki, N. Székely, N. Tóth: "Neural Network Based Microcalcification Detection in a Mammographic CAD System", *International Scientific Journal of Computing*, Vol. 3, Issue 3, 2004.

6.2 Magyar folyóiratcikkek angol nyelven

- [p 4] N. Tóth, B. Pataki: "A Proposed Method to Handle Classification Uncertainty Using Decision Trees", *Production Systems and Information Engineering*, Vol. 3, 2006, pp. 21-38.

6.3 Magyar folyóiratcikkek magyar nyelven

- [p 5] N. Tóth, B. Pataki: "Foltok detektálása mammogramokon textúra-analízis segítségével" = "Detecting Lesions in Mammograms Using Texture Analysis", *HÍRADÁSTECHNIKA*, Budapest, 2005 April, pp.: 22-24 (in Hungarian).
- [p 6] N. Tóth, B. Pataki: "Mammogramok textúra analízise" = "Texture Analysis of Mammograms", *INFORMATIKA*, Budapest, 2003 November, pp. 32-39 (in Hungarian).
- [p 7] N. Tóth, B. Pataki: "Textúra analízis" = "Texture Analysis", *ELEKTROnet*, Budapest, 2003/3, pp. 51-53 (in Hungarian).

6.4 Nemzetközi konferenciák

- [p 8] N. Tóth, B. Pataki: "On Classification Confidence and Ranking Using Decision Trees", *Proceedings of the 11th International Conference on Intelligent Engineering Systems (INES2007)*, Budapest, Hungary, June 29 - July 1, 2007, pp. 133-138.
- [p 9] N. Tóth, G. Takács, B. Pataki: "Mass Detection in Mammograms Combining Two Methods", in *Proceedings of the 3rd European Medical & Biological Engineering Conference (EMBE'05)*, Prague, Czech Republic, November 20-25, 2005.
- [p 10] N. Székely, N. Tóth, L. Lasztovicza, B. Pataki: "Combining Methods For Mass Detection In Mammograms", *Proceedings of 17th Biennial International EURASIP Conference BIOSIGNAL 2004*, Brno, Czech Republic, 23-25 May 2004, pp. 281-283.

- [p 11] L. Lasztovicza, N. Tóth, N. Székely, B. Pataki: "*Hybrid Microcalcification Detection In Mammograms*", *Proceedings of 17th Biennial International EURASIP Conference BIOSIGNAL 2004*, Brno, Czech Republic, 23-25 May 2004, pp. 287-289.
- [p 12] N. Székely, N. Tóth, B. Pataki: "A Hybrid System for Detecting Masses in Mammographic Images", *Proceedings of IMTC/04, 21th IEEE Instrumentation and Measurement Technology Conference*, Como, Italy, 18-20 May 2004, pp. 2065-2070.
- [p 13] László Lasztovicza, Béla Pataki, Nóra Székely, Norbert Tóth: "Neural Network Based Microcalcification Detection in a Mammographic CAD System", *Proceedings of the Second IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2003)*, Lviv, Ukraine, 8-10 September, 2003, pp. 319-323.

6.5 Technical Reportok

- [p 14] N. Tóth, B. Pataki: "*Extending the Decision Tree Framework to Handle Classification Certainty*", Budapest University of Technology and Economics, Department of Measurement and Information Systems, Hungary, 2007.

6.6 Hazai konferenciák

- [p 15] N. Tóth, B. Pataki: "Decision Trees with Uncertainty", *Proceedings of the 13th Mini-Symposium*, Budapest, Hungary, 6-7 February 2006, pp. 42-43.
- [p 16] N. Tóth, B. Pataki: "Detecting Masses in Mammograms Using Texture Analysis", *Proceedings of the 12th Mini-Symposium*, Budapest, Hungary, 8-9 February 2005, pp. 22-23.
- [p 17] N. Tóth, B. Pataki: "Texture Analysis of Mammograms", *Proceedings of the 11th Mini-Symposium*, Budapest, Hungary, 3-4 February 2004, pp. 52-53.

7 Irodalomjegyzék

- [1] E. J. Feuer, L. M. Wun, C. C. Boring, W. D. Flanders, M. J. Timmel, T. Tong: "The Lifetime Risk of Developing Breast Cancer", *Journal of the National Cancer Institute*, 1993, 85(11), pp. 892-897.
- [2] G. Horváth, B. Pataki, J. Valyon, Zs. Dömötöri, N. Székely, N. Toth, G. Takács: "An Intelligent Decision Support System for Screening Mammography", *Proc. of the 3rd European Medical and Biological Engineering Conference EMBEC'05*, Prague, Czech Republic, Nov. 2005.
- [3] S. Boyd, L. Vandenberghe: *Convex Optimization*, Cambridge University Press, 2006.
- [4] P. C. Mahalanobis: "On the generalised distance in statistics", *Proceedings of the National Institute of Science of India*, 12 (1936) 49-55.
- [5] B. W. Silverman: *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
- [6] R. O. Duda, P. E. Hart, D. G. Stork: *Pattern Classification*, John Wiley & Sons, 2001.
- [7] D. Harde: *Non Asymptotic Binomial Confidence Intervals*, Statistics Research Associates, Wellington, web: <http://www.statsresearch.co.nz/pdf/confint.pdf>.
- [8] D. Harde: *Non Asymptotic Binomial Confidence Intervals*, Statistics Research Associates, Wellington, web: <http://www.statsresearch.co.nz/pdf/confint.pdf>.
- [9] J. Kittler, M. Hatef, R. P.W. Duin: "On Combining Classifiers", *IEEE Transactions on Pattern Analysis and Machine Learning*, vol. 20, no. 3, March 1998.
- [10] M. I. Jordan, R. A. Jacobs: "Hierarchical Mixtures of Experts and the EM Algorithm", *Neural Computation*, 6, pp. 181-214, 1994.
- [11] L. Breiman: "Bagging Predictors", *Machine Learning*, vol. 24, 2, pp. 123-140, 1996.
- [12] Á. Zarándy, T. Roska, Gy. Liszka, J. Hegyesi, L. Kék, and Cs. Rekeczky: "Design of Analogic CNN Algorithms for Mammogram Analysis", *Proceedings of the IEEE International Workshop on Cellular Neural Networks and their Applications (CNNA'94)*, pp. 255-260, Rome, 1994.
- [13] Kék László, Liszka György Dr., Petrányi Ágota Dr., Zarándy Ákos Dr., Bölöni László: "Mammographiás szűrőállomásokhoz telepített analogikai munkahely adatkezelése", *Magyar Onkológia* 42., 109-120, 1998.
- [14] DDSM: Digital Database for Screening Mammography, <http://marathon.csee.usf.edu/Mammography/Database.html>