BASTIAN: Incorporació de la teoria dels Rough Sets dins del Raonament Basat en Casos

Maria Salamó[†], Elisabet Golobardes[†]

[†]Grup de Recerca en Sistemes Intel·ligents
Departament d'Inmformàtica
Enginyeria i Arquitectura La Salle
Universitat Ramon Llull
Passeig Bonanova, 8 08022-Barcelona
{mariasal,elisabet}@salleURL.edu

Resum

Aquest article proposa com incorporar la teoria dels Rough Sets com a mètode de ponderació dins d'un sistema classificador basat en casos. Aquest sistema híbrid s'ha implementat sobre la plataforma BASTIAN (case-**BA**sed **S**ys**T**em **I**n cl**A**ssificatio**N**), la qual incorpora ambdues tècniques.
Els principals objectius d'aquest article són: presentar el sistema BASTIAN; i analitzar aquesta proposta utilitzant diferents dominis, extrets del repositori UCI.

Paraules clau: Raonament Basat en Casos, Aprenentatge Artificial, Diagnosi, Descobriment de Coneixement.

1 Introducció

El nostre principal objectiu és desenvolupar, avaluar i millorar els sistemes classificadors. En aquest article presentem un sistema classificador híbrid basat en Raonament Basat en Casos i Rough Sets. La plataforma BASTIAN és un sistema de Raonament Basat en Casos que incorpora els Rough Sets per a millorar la predicció del sistema. Rough Sets és utilitzat en el nostre sistema com a mètode de ponderació per a seleccionar la millor rellevància dels atributs en el domini.

El Raonament Basat en Casos (CBR) [16] s'ha utilitzat àmpliament en una gran varietat de camps i aplicacions. En el nostre cas, utilitzem el CBR com un sistema automàtic de classificació [4,21].

Rough Sets és una tècnica de Data Mining. La naturalesa d'aquesta tècnica fa que es pugui aplicar per a reduir el coneixement, extraure dependències, raonar, reconeixement de patrons, etc.

Les principals àrees de recerca en Rough Sets són:

1. El tractament de coneixement incomplet.
2. La manipulació de peces inconsistents d'informació.
3. La manipulació de diferents nivells de representació, des de universos refinats fins a universos gruixuts (coarser) i al revés.

Aquest article es divideix en les següents seccions. En primer lloc, en una descripció del la plataforma BASTIAN (veure secció 2). A continuació es presenta la teoria de Rough Sets com a mètode de ponderació per a un sistema classificador Basat en Casos. Les seccions 4 i 5 exposen el banc de proves utilitzat i els resultats obtinguts, respectivament. Finalment, l'última secció presenta les conclusions i el treball futur.

2 Descripció de la plataforma BASTIAN

La plataforma BASTIAN és un sistema de Raonament Basat en Casos utilitzada per a problemes de classificació. El raonament basat en casos integra dos característiques ben diferenciades: d'una banda, l'aprenentatge i de l'altre la resolució de problemes.

El CBR utilitza una filosofia similar a la que, de vegades, utilitzem els humans: intenta resoldre nous casos (exemples) d'un problema utilitzant casos resolts anteriorment. El procés de resolució de nous casos contribueix amb nova informació dins del sistema. Aquesta nova informació podrà ser utilitzada posteriorment per a resoldre casos futurs. El procés bàsic del CBR es pot descriure en 4 fases [1,11]. La figura 1, mostra la descripció d'aquestes 4 fases, definides com:

- *Retrieval (Recuperació)*: aquesta fase selecciona dins de la memòria de casos, el cas o casos més similars al cas nou que volem resoldre.

- *Reuse (Adaptació)*: el sistema intenta adaptar el cas o casos recuperats per a resoldre el cas nou.

- *Revise (Revisió)*: en aquesta fase es revisa la solució proposada en la fase anterior.

- *Retain (Emmagatzematge)*: el sistema emmagatzema tota aquella informació rellevant obtinguda al resoldre el nou cas.

En un sistema Classificador Basat en Casos, és possible simplificar la fase d'adaptació classificant el nou cas amb la mateixa classe que el cas més similar recuperat de la memòria de casos.

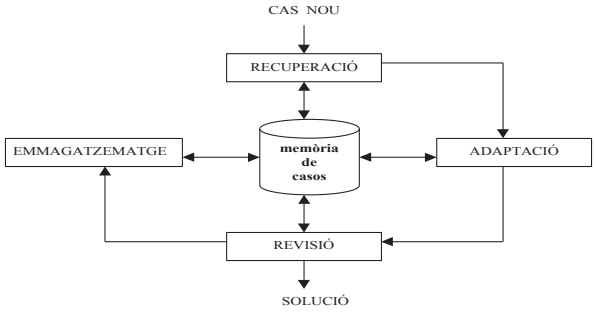

Figura 1: Cicle del Raonament Basat en Casos

BASTIAN és una extensió del sistema CaB-CS (**Ca**se-**B**ased **C**lassifier **S**ystem) [9,7,6]. BASTIAN permet la configuració de diferents variants de CBR per a ser provades per un usuari. Exactament, la variant presentada en aquest article està enfocada en dos fases del cicle de CBR: la fase de recuperació i d'emmagatzematge, i també en l'organització de la memòria de casos. BASTIAN ha esta desenvolupat en llenguatge JAVA i el sistema s'està millorant amb noves funcionalitats.

2.1 Estructura General

L'estructura general del BASTIAN, veure figura 2, manté les quatre fases descrites en [1]. El sistema afegeix una fase inicial anomenada *StartupInterface*, no inclosa en el cicle del Raonament Basat en Casos, que prepara la posta en marxa i inicialitzacions del sistema.

Totes les funcionalitats del sistema han estat desenvolupades per a treballar separadament i independentment, però a la vegada en cooperació entre elles. Cadascuna de les funcionalitats descrites, en l'estructura general, conté una descripció del comportament general que ha d'assolir. El nostre principal objectiu ha estat obtenir una estructura general que pugui canviar dinàmicament depenent del tipus de CBR que es vol desenvolupar.

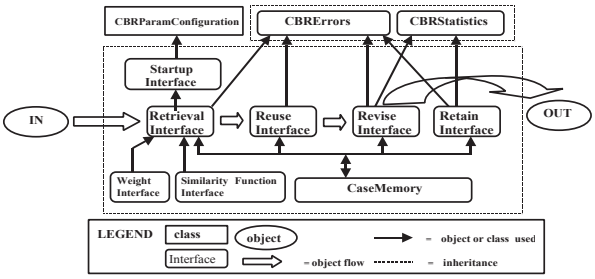

Figura 2: Estructura General del BASTIAN.

Les principals funcionalitats són:

- *CBRParamConfiguration*: ens modifica, guarda i recupera la configuració del sistema. El sistema de configuració és independent de la plataforma. Així permet la configuració dinàmica de tota la plataforma en qualsevol moment.

- *CBRErrors*: és el controlador d'errors de la plataforma. S'encarrega de detectar i emmagatzemar tots els errors durant l'execució.

- *CBRStatistics*: nucli inicial d'estadístiques de tot el sistema durant l'execució. El format de les estadístiques es realitza en format EXCEL, LATEX i EPS.

- *CaseMemory*: desenvolupa diferents tipus d'organitzacions de la memòria de casos.

- *SimilarityFunctionInterface*: defineix les operacions bàsiques a desenvolupar per les funcions de similitud. Permet, a més a més, modificar la funció de similitud durant una execució.

- *WeightingInterface*: conté les principals característiques que s'han d'acomplir per a calcular la rellevància dels atributs dins del sistema classificador basat en casos. Està relacionada amb *RetrievalInterface* i *SimilarityFunctionInterface*.

- {Retrieval, Reuse, Revise, Retain}Interface: són les quatre fases del cicle de CBR. Aquestes interficies descriuen el comportament de cada fase.

El nucli d'un sistema de Raonament Basat en Casos és la fase de recuperació (fase 1). Aquesta fase recupera el cas o casos més similars al cas nou que es vol resoldre. La similitud entre dos casos es calcula a partir de diferents funcions de similitud. El nostre objectiu és millorar l'aproximació d'aquestes funcions de similitud utilitzant mètodes de ponderació d'atributs que calculin la rellevància de cadascun d'ells.

2.2 Funcions de similitud

Donat l'objectiu d'aquest article, hem utilitzat les funcions de similitud, basades en el concepte de distància, que s'han introduït dins del BASTIAN. La més utilitzada és *Nearest Neighbour Algorithm*, la qual calcula la similitud entre dos casos usant una mesura de similitud global [2,3]. La implementació pràctica (utilitzada en el nostre sistema) d'aquesta funció està basada en la *mètrica de Minkowsky* [6,12], també utilitzem la distància de *Clark* i la distància del *cosinus*.

Mètrica de Minkowsky

La mètrica de Minkowsky es defineix com:

$$Sim(Cas\_x, Cas\_y) = \sqrt[r]{\sum_{i=1}^{F} w_i \times |x_i - y_i|^r} \quad (1)$$

On Cas_x i Cas_y són els dos casos dels que estem calculant la similitud; F és el nombre d'atributs que descriuen un cas; x_i, y_i representen el valor i-èssim pel cas x i y, respectivament; i w_i és el pes de l'atribut i-èssim.

En aquest estudi hem utilitzant la mètrica de Minkowsky amb tres valors diferents del paràmetre *r*: distància de *Hamming* per a r=1, distància *Euclidiana* per a *r=2*, i distància *cúbica* per a *r=3*.

Distància de Clark

La distància de Clark es defineix com:

$$Sim(Cas\_x, Cas\_y) = \sqrt[2]{\sum_{i=1}^{F} w_i \times \frac{|x_i - y_i|^2}{|x_i + y_i|^2}} \quad (2)$$

On *F* representa el nombre total d'atributs que descriuen un cas; x_i i y_i representen el valor de l'atribut i-èssim pels casos *x* i *y*, respectivament; i w_i és el pes de l'atribut i-èssim.

Distància del cosinus

La distància del cosinus està basada en les propietats dels vectors en un espai euclidià. Aquesta distància mesura l'angle del cosinus en un espai *n* dimensional. La distància es defineix com:

$$Sim(Cas\_x, Cas\_y) = w_i \times \frac{\sum_{i=1}^{F} (x_i \times y_i)}{\sqrt[2]{(\sum_{i=1}^{f} x_i^2) \times (\sum^{F} y_i^2)}} \quad (3)$$

On *F* representa el nombre total d'atributs que descriuen un cas; x_i i y_i representen el valor de l'atribut i-èssim pels casos *x* i *y*, respectivament; i w_i és el pes de l'atribut i-èssim.

2.3 Representació de la Memòria de Casos

La representació de la memòria de casos està descrita en la figura 3. Tal i com es pot observar, existeixen tres estructures per a utilitzar en la plataforma BASTIAN: la primera d'elles, és una llista (*list*); la segona és un vector (*SingleList*) i l'última és un arbre (*tree*).

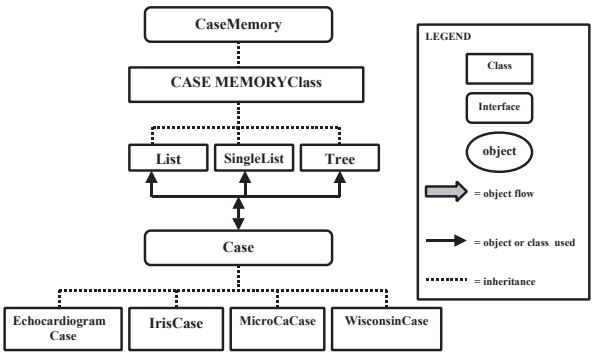

Figura 3: Estructura de la memòria de casos en BASTIAN

La memòria de casos utilitzada en els nostres experiments ha estat la primera, una llista de casos. La segona part de la figura 3 mostra els problemes que hem utilitzat en el nostre estudi.

La representació utilitzada en cada exemple està basada en una representació atribut-valor, veure equació 4.

$$Cas = \{a_0, a_1, a_2, ..., a_n, CLASS\} \quad (4)$$

On a_i és el valor per l'atribut i-èssim, i *CLASS* és la classe a la que pertany el cas.

2.4 Polítiques de Retain

Per tal de decidir si un cas és suficientment representatiu per ser emmagatzemat en la memòria de casos, utilitzem tres polítiques de retain diferents, mirar figura 4.

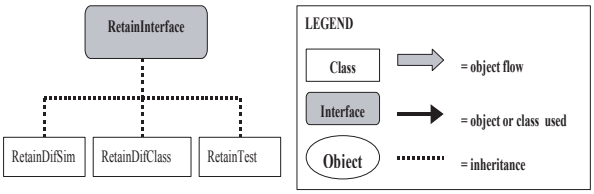

Figura 4: Estructura fase Retain en la plataforma BASTIAN

- **Mode Test**, en aquest mode el sistema no emmagatzema cap nou cas dins de la memòria de casos. Aquest criteri s'ha escollit per dues raons. La primera d'elles és que els resultats obtinguts amb aquest mode poden ser comparats, amb igualtat de condicions, amb altres tècniques d'aprenentatge artificial que no inclouen l'aprenentatge quan estan resolent nous problemes. L'altre raó és que ens permet avaluar el *corpus* inicial de la memòria de casos.
- **Mode DifSim**, aquesta política emmagatzema un nou cas si la seva similitud amb el cas recuperat és diferent de cero. En altres paraules, el cas nou s'emmagatzema si no és idèntic al cas que s'ha recuperat.
- **Mode DifClass**, és una política a mig camí entre les dues anteriors. El sistema emmagatzemarà un nou cas si aquest no s'ha pogut classificar correctament . Altrament, no s'emmagatzemarà.

El sistema ens permet, a més a més, entrenar la memòria de casos inicials per tal d'emmagatzemar només els casos més representatius.

3 Rellevància dels atributs

BASTIAN inclou tres variants com a mètodes per a ponderar la rellevància dels atributs. El primer mètode és la Correlació Mostral [9]; el segon d'ells és l'Entropia de Shannon [13]; i finalment, el mètode basat en la teoria dels Rough Sets [18]. L'objectiu d'aquest article és explicar la integració de l'últim mètode de ponderació dins de la plataforma BASTIAN. Els Rough Sets dins del sistema BASTIAN es poden aplicar de dues maneres diferents:

- **Estàticament**, es calcula la ponderació d'atributs només sobre la memòria de casos inicials. Aquest article es centrarà en aquesta proposta.
- **Dinàmicament**, la rellevància dels atributs es calcula sobre la memòria de casos inicials, i també cada vegada que s'emmagatzema un nou cas sobre la memòria de casos. A aquest sistema de ponderació li direm incremental.

Aquesta secció l'hem dividit en una introducció de la teoria dels Rough Sets, alguns conceptes bàsics sobre la teoría i, per acabar, amb l'explicació de la incorporació dels Rough Sets dins del Sistema de Raonament Basat en Casos.

3.1 La teoria dels Rough Sets

La teoria dels Rough Sets va ser introduïda per Zdzislaw Pawlak l'any 1982 [14, 15, 16]. La idea bàsica dels Rough Sets consisteix en aproximar un conjunt per un parell de conjunts, anomenats l'aproximació inferior i superior d'aquest conjunt. Concretament, aquestes aproximacions són les operacions de clausura i pertinença per a una certa topologia generada amb les dades existents amb els elements d'aquest conjunt.

La naturalesa dels Rough Sets, fa que aquesta teoria sigui útil per a reduir coneixement, extraure dependències del coneixement, raonar, reconeixement de patrons, etc.

En el nostre cas, utilitzem la teoria del Rough Sets per a reduir i extraure les dependències del nostre coneixement. Aquestes dependències són els

fonaments per a calcular la rellevància de cadascun dels atributs dins del Raonament Basat en Casos.

3.2 Rough Sets dins del Raonament Basat en Casos

Com incorporem la teoria dels Rough Sets dins del nostre sistema classificador basat en casos?

En primer lloc, en aquest article incorporem alguns conceptes per a explicar com obtenir les dependències que estem buscant dins del nostre coneixement per tal d'extraure el millor conjunt de pesos.

Conceptes bàsic i definicions

A partir del nostre **Univers (U)** (conjunt finit no buit d'objectes que descriuen el nostre problema, la memòria de casos), calculem els **conceptes** (objectes o casos) que formen les particions del nostre univers. La unió de tots els conceptes formen tot l'univers. Utilitzant tots els conceptes podem descriure les **relacions d'equivalència** sobre el nostre univers (U). Una relació d'equivalència és el conjunt d'atributs que descriuen un concepte. **U/R** són les famílies de totes les classes d'equivalència de (R).

L'univers i les relacions formen la **base de coneixement (KB)**, definida com **KB=<U, $\hat{\mathbf{R}}$ >**. On $\hat{\mathbf{R}}$ és la família de relacions d'equivalència sobre U. Cada relació sobre l'univers és un concepte elemental de la base de coneixement.

Tots els conceptes estan formats per un conjunt de relacions d'equivalència que els descriuen. Així, nosaltres buscarem el mínim conjunt de relacions d'equivalències que defineixen el mateix concepte que el conjunt inicial.

Definició 1 (Relacions d'indiscernibilitat)

Es defineix com IND($\hat{\boldsymbol{P}}$)= $\cap$ $\hat{\mathbf{R}}$ on $\hat{\boldsymbol{P}} \subseteq \hat{\mathbf{R}}$. Si fem la lectura podem dir que la relació d'indiscernibilitat és la intersecció de totes les propietats de P. La indiscernibilitat dóna la informació més refinada d'un concepte. Dóna la informació de totes les relacions d'equivalència que pertanyen a $\hat{\boldsymbol{P}}$.

Exemple 3.1

Suposem que tenim un univers amb 8 objectes o casos, U = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8), amb la següent família de relacions d'equivalència $\hat{\mathbf{R}}$ ={P, Q, S}. On P són colors (verd, blau, vermell i groc); Q és mida (petit , gran, mig); i S són formes (quadrat, rectangular, circular, triangular).

U/P=((x_1, x_4, x_5), (x_2, x_8), (x_3), (x_6, x_7))
U/Q=((x_1,x_3, x_5), (x_6), (x_2, x_4, x_7, x_8))
U/S=((x_1, x_5), (x_6), (x_2, x_7, x_8), (x_3, x_4))

Tal i com es pot veure, cada relació d'indiscernibilitat divideix l'Univers de forma diferent.

Definició 2 (Coneixement bàsic)

El coneixement bàsic és la família de totes les classes d'equivalència de les relacions d'equivalència relacionades amb IND($\hat{\boldsymbol{P}}$). El coneixement bàsic mostra tot el coneixement associat amb la família de relacions d'equivalència de P.

Definició 3 (Categories P-bàsiques)

Les categories P-bàsiques són les propietats bàsiques de l'univers, les quals poden ser expressades usant coneixement de P. Són els fonaments del nostre coneixement.

Suposant una base de coneixement **KB=<U, $\hat{\mathbf{R}}$ >.**
IND(K)=IND($\hat{\boldsymbol{P}}$): $0 \neq \hat{\boldsymbol{P}} \subseteq \hat{\mathbf{R}}$ és la família de totes les relacions d'equivalències definides en K.

Definició 4 (equivalència i generalització)

Tenint K i K' bases de coneixement:

- Si IND(K) = IND(K'), això significa que k i K' són equivalents.
- Si IND(K) $\subset$ IND(K') llavors la base de coneixement K és més fina que K', per tant K' és una generalització de K.

Rough Sets

Si tenim X $\subseteq$ U i R com a relació d'equivalència. Direm que:

- X és *R-definible* si X és la unió d'alguna categoria R-bàsica; altrament direm que X és *R-indefinible*.
- Els conjunts *R-definible* són aquells subconjunts del Univers que poden ser exactament definits en la base de coneixement K, mentre que *R-indefinible* són aquells que no es poden definir en la base de coneixement.
- El conjunt *R-indefinible* es pot anomenar també *R-Rough*.

- El conjunt $X \subseteq U$ s'anomena *exacte* en K si existeix $R \in IND(K)$ tal que X és R-exacte, i X s'anomena Rough en K, si X és R-Rough per qualsevol $R \in IND(K)$.

Aproximacions dels conjunts. Aquesta és la idea principal dels Rough Sets, aproximar un conjunt a partir d'altres conjunts. Les següents definicions mostren aquesta idea.

Partim d'una base de coneixement **KB=<U, $\hat{\mathbf{R}}$ >**, amb un conjunt $X \subseteq U$ i una relació d'equivalència $R \subseteq IND(K)$, es poden associar dos conjunts:

- Aproximació inferior (lower)
- Aproximació superior (upper)

Definició 5 (Aproximació inferior)

L'aproximació inferior es defineix com : $\underline{\mathbf{R}}\, X = \cup \{Y \in U/R: Y \subseteq X\}$. L'aproximació inferior és el conjunt de tots els elements de U que amb certesa poden ser classificats com elements de X, utilitzant el coneixement de R.

Definició 6 (Aproximació superior)

L'aproximació superior es defineix com: $\overline{\mathbf{R}}\, X = \cup \{Y \in U/R: X \cap Y \neq \varnothing\}$. L'aproximació superior és el conjunt de tots els elements de U que possiblement poden ser classificats correctament com a elements de X, utilitzant el coneixement R.

Definició 7 (Frontera)

$\overline{\mathbf{R}}\, X - \underline{\mathbf{R}}\, X$ és la frontera BNr(X). La frontera és el conjunt d'elements que no poden ser classificats com a lements de X o elements de $\neg X$ tenint el coneixement R.

Reducts and Core. Intuïtivament, un reducte (reduct) de coneixement és la part essencial d'aquest coneixement, el qual és suficient per a definir tots els conceptes del coneixement, mentre que un nucli (core) és la part més important d'aquest coneixement.

Partint de $\hat{\mathbf{R}}$ com a família de relacions d'equivalències i $R \in \hat{\mathbf{R}}$. Podem dir que:

- R és indispensable si $IND(\hat{\mathbf{R}}) \neq IND(\hat{\mathbf{R}} - R)$; altrament és dispensable.
- La família $\hat{\mathbf{R}}$ és independent si cada $R \in \hat{\mathbf{R}}$ és indispensable en R; altrament és dependent.

Definició 8 (Reduct)

$\hat{\mathbf{Q}} \in \hat{\mathbf{R}}$ és un reducte de $\hat{\mathbf{R}}$ si:

1. $\hat{\mathbf{Q}}$ és independent
2. $IND(\hat{\mathbf{Q}}) = IND(\hat{\mathbf{R}})$. Utilitzant Q és possible aproximar el mateix que utilitzant R.

Definició 9 (Core)

El conjunt de totes les relacions indispensables en R s'anomena *core* de R, i es defineix com:

$$\mathbf{CORE}(\hat{\mathbf{R}}) = \bigcap \mathbf{RED}(\hat{\mathbf{R}})$$

on $RED(\hat{\mathbf{R}})$ és la família de reductes de R.

Exemple 3.2

Continuem amb l'exemple 3.1 per a buscar els reductes i el core de la nostra base de coneixement. Les classes d'equivalència eren:

$U/P=((x_1, x_4, x_5), (x_2, x_8), (x_3), (x_6, x_7))$
$U/Q=((x_1,x_3, x_5), (x_6), (x_2, x_4, x_7, x_8))$
$U/S=((x_1, x_5), (x_6), (x_2, x_7, x_8), (x_3, x_4))$

Així la relació IND(R) té la següent classe d'equivalència:
$U/IND(\hat{\mathbf{R}}) = ((x_1, x_5), (x_2, x_8), (x_3),(x_4),(x_6), (x_7))$

La relació P és indispensable en R, ja que:

$U/IND(\hat{\mathbf{R}} - P) = ((x_1, x_5), (x_2, x_7, x_8), (x_3) ,(x_4),(x_6)) \neq U/IND(\hat{\mathbf{R}})$

En canvi, les relacions Q i S són dispensables ja que aproximen el mateix que R.
$U/IND(\hat{\mathbf{R}} - Q) = ((x_1, x_5), (x_2, x_8), (x_3),(x_4),(x_6), (x_7)) = U/IND(\hat{\mathbf{R}})$

$U/IND(\hat{\mathbf{R}} - S) = ((x_1, x_5), (x_2, x_8), (x_3),(x_4),(x_6), (x_7)) = U/IND(\hat{\mathbf{R}})$

Així doncs, la classificació definida pel conjunt de tres relacions d'equivalència P, Q, i S és la mateixa que la classificació definida per P i Q or P i S.

Per tant, els reductes i el core són:

RED($\hat{\mathbf{R}}$) = ((P,Q), (P,S))
CORE ($\hat{\mathbf{R}}$) = (P)

Com introduir els Rough Sets dins del Raonament basat en casos?

Utilitzarem la informació dels reductes i el core per a ponderar la rellevància dels nostres atributs en el sistema. Un atribut que no apareix en els reductes tindrà una rellevància de valor 0.0, mentre que un atribut que apareix en tots els reductes, per tant forma el core tindrà una rellevància de 1.0. La resta d'atributs tindran un valor proporcional a la seva aparició en els reductes. Aquest sistema de ponderació basat en la proporcionalitat d'aparició en els reductes i core és que utilitzarem dins del nostre sistema classificador basat en casos.

La figura 5 mostra, a grans trets, quins són els passos que realitzem per a calcular la ponderació d'atributs, a partir de la teoria dels Rough Sets, i utilitzar-la en el nostre sistema classificador. Aquest dibuix, mostra com es realitza aquesta incorporació dins del nostre sistema.
Els procés de ponderació està dividit en tres passos. El primer pas és discretitzar les dades, és necessari per tal d'utilitzar la teoria dels Rough Sets; el segon pas és la cerca dels reductes i el core dins de la base de coneixement que tenim fins aquest moment, utilitzant la teoria dels Rough Sets; i finalment, com a darrer pas, utilitzem la informació dels reductes i el core per a treure el valor de ponderació proporcional dels nostres atributs.

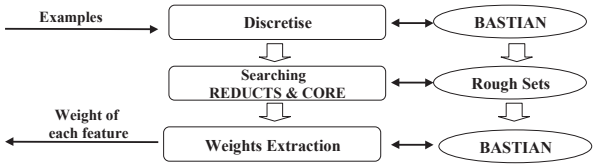

Figura 5. Meta procés dels Rough Sets.

La teoria dels Rough Sets s'ha introduït com a mètode de ponderació en dos fases del cicle modificat de CBR. La primera fase és la fase d'*start-up* i la segona és la fase de *retain*. La fase d'start-up calcula els pesos amb la memòria de casos inicial, aquests pesos seran utilitzats per la fase de *retrieval* posteriorment. La fase d'emmagatzemament (*retain*) calcula els pesos de la memòria de casos, sempre i quan s'hagin emmagatzemat nous casos dins de la memòria de casos. La codificació del sistema de Rough Sets ha estat realitzada utilitzant una llibreria pública *Rough Sets Library* [8].

4 Banc de proves

La experimentació s'ha realitzat usant 3 bancs de proves del UCI repository (echocardiogram, iris, breast cancer Wisconsin) i un conjunt de dades del nostre banc de proves (mammogram problem). Les Taules 1 i 2 mostren les característiques de cada banc de proves. E problema de les mamografies consisteix en la detecció de cancer de mama a partir de les microcalcificacions que es pot trobar en una mamografia digital. Una microcalcificació (μCa) normalment apareix en les mamografies com un punt petit, brillant i de formes diferents. Així doncs, la seva anàlisis i caracterització es realitza a través de l'extracció de característiques i descriptors de visibilitat, utilitzant diferents tècniques de processament d'imatges [19]. Cadascun dels exemples (mamografies) conté la descripció de diferents μCa. Cada μCa està descrita per 23 atributs reals. En altres paraules, una mamografia està representada per una matriu de *m* files x *23* columnes, on *m* és variables segons el nombre de microcalcificacions presents en la mamografia. Aquest banc de proves conté 216 mamografies.

Domini	**Referència**
Echocardiogram	E
Iris	I
Breast Cancer Wisconsin	BC
Mammogram	M

Taula 1. Banc de proves utilitzat.

Ref	Exem-ples	Atri-buts	Cla-sses	Valors desconeguts	Incon-sistent
E	132	9	2	132	Si
I	150	4	3	0	No
BC	699	9	2	9	Si
M	216	23	2	0	Si

Taula 2. Característiques dels bancs de proves.

Els diferents problemes han estat dividits en dos conjunts: el conjunt d'entrenament i el conjunt de test. El conjunt d'entrenament s'usa per a entrenar el sistema, mentre que el segon conjunt és l'utilitzat per a testejar el sistema. El conjunt d'entrenament i de test s'ha generat utilitzant diferents proporcions

d'exemples: 10% d'entrenament i 90% de test, 20% d'entrenament i 80% de test, fins arribar a la proporció 90% d'entrenament i 10% de test.

Hem provat cada conjunt de dades utilitzant les següents polítiques:

- **Funcions de similitud**: Mètrica de Minkowski (distància de Hamming, Euclidiana i Cúbica), distància de Clark i Cosinus.
- **Polítiques de Retain**: DifSim, DifClass i Test.
- **Entrenament conjunt inicial de dades**: entrenament de la memòria de casos inicial i manteniment de la memòria de casos inicial.
- **Exemples**: tenim 9 proporcions de cada conjunt de dades i 10 versions per a cada proporció.

Per a cada conjunt de dades fem un total de 2700 execucions.

5 Resultats

En aquesta secció es presenten els resultats més importants obtinguts de les execucions anteriors. La taula 3 presenta els resultats obtinguts durant l'execució 90% d'entrenament 10% de test. La primera columna mostra els resultats obtinguts de l'execució de la plataforma BASTIAN sense utilitzar ponderació d'atributs, la segona columna mostra els resultats per al mètode de ponderació basat en la teoria dels Rough Sets.

Ref	Not Weighting (¬W)	Rough Sets weights (PRS)
E	78.57%	78.57%
I	100%	100%
BC	98.71%	98.71%
M	77.27%	81.81%

Taula 3. Resultats màxims per a cada conjunt de dades

Aquesta proporció s'ha escollit per la rellevància dels resultats obtinguts. Val a dir que els resultats presentats en aquesta taula són els millors resultats obtinguts en una execució.

Els resultats presentats obtenen una bona mesura de predicció. Volem fer èmfasis, en que els resultats màxims que apareixen utilitzant el mètode de ponderació basat en la teoria de Rough Sets, apareixen més freqüentment que quan no s'usa mètode de ponderació.

La figura 6 mostra els resultats obtinguts per totes les proporcions d'entrenament en el problema de les mamografies. Tal i com es pot veure, el mètode de ponderació necessita un volum de dades d'entrenament més elevat per tal de desenvolupar una bona rellevància dels atributs. En canvi, la predicció del sistema augmenta quan es disposa de suficient informació en la memòria de casos per tal de desenvolupar un bon criteri de ponderació. A més a més, el sistema decrementa la desviació estàndard quan usa el mètode de ponderació.

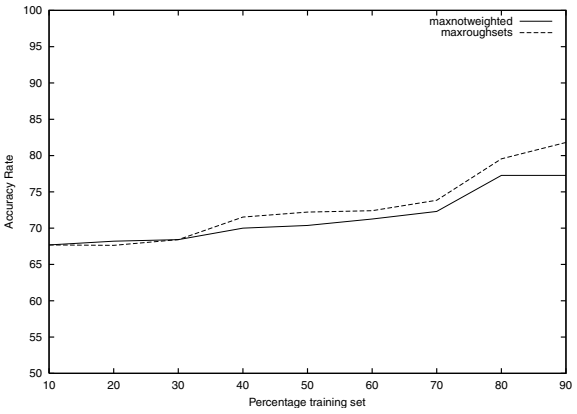

Figura 6. Resultats màxims per al problema de les mamografies.

També es pot observar que és molt important seleccionar una bona memòria de casos inicial per tal d'aconseguir millors resultats. Així, els millors resultats s'han obtingut quan s'entrenava a priori la memòria de casos inicial. La memòria de casos s'ha decrementat utilitzant aquesta política, deixant d'aquesta manera, els casos més rellevants de la memòria inicial.

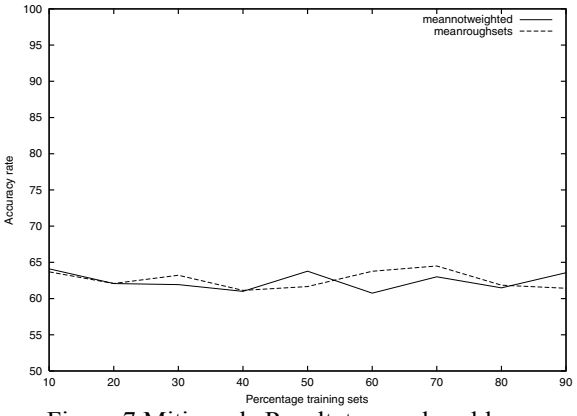

Figura 7.Mitjana de Resultats per al problema echocardiogram.

La taula 4 mostra els resultats obtinguts en diferents proporcions d'entrenament per al problema Iris. Els resultats presentats són el màxim i la mitjana per a cada proporció, usant i no usant mètode de ponderació. Tal i com es pot observar, existeixen poques diferències entre el sistema híbrid amb Rough Sets i el sistema classificador basat en casos pur. Els resultasts denoten, a més a més, que és molt important el nombre de casos que s'inclouen en la memòria de casos per tal d'aconseguir una bona ponderació dels atributs.

És important remarcar que el percentatge de classificació depèn de la mida de la memòria de casos. Aquest fet es pot observar en tots els problemes que s'han provat.

La figura 7 mostra la mitjana de resultats obtinguts per al problema echocardiogram en totes les proporcions d'entrenament. També reafirma la conclusió anterior, que és molt important el nombre de casos que tenim dins de la memòria de casos i també que els resultats depenen del nombre d'atributs sense valor (missing values) que tenim en el sistema.

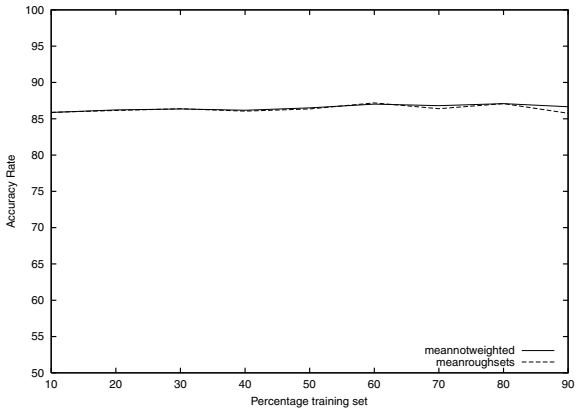

Figura 8. Mitjana problema breast cancer Wisconsin

Els resultats obtinguts per al problema breast cancer Wisconsin es poden trobar en la figura 8. Els resultats són molt similars als presentats anteriorment, això és degut al gran volum de dades que existeix per aquest problema i al nombre de missing values que té.

Finalment, és important remarcar que tots els problemes s'han discretitzat utilitzant el mateix nombre d'intervals. Aquest paràmetre hauria de ser modificat segons els límits inferiors i superiors de cada atribut. Aquesta metodologia de discretització de les dades pot influenciar àmpliament els resultats obtinguts.

6 Conclusions i treball futur

Aquest article ha presentat la introducció de la teoria dels Rough Sets com a mètode de ponderació d'atributs dins d'un sistema classificador basat en casos. Aquest treball tracta principalment dos punts: la proposta d'una plataforma que incorpora la teoria dels Rough Sets dins d'un sistema classificador basat en casos (anomenat BASTIAN) i la millora de la ponderació d'atributs a partir de la teoria dels Rough Sets.

El nostre mètode de ponderació basat en la teoria de Rough Sets s'ha avaluat amb diferents problemes del UCI repository. Entre les conclusions dels nostres resultats, podem dir que la més important ha estat observar com el bon funcionament en predicció del mètode de ponderació depèn àmpliament del volum de dades que tenim en la memòria de casos i també de si aquestes dades són prou representatives o no. També és important dir que el sistema només decrementa la seva predicció en el cas de tenir una proporció d'entrenament del 10%. El sistema de ponderació basat en la teoria dels Rough Sets ajuda a equilibrar els resultats en el sistema, provocant que el sistema no tingui una desviació molt àmplia.

El nostre treball futur en aquesta àrea està enfocat a millorar la predicció utilitzant diferents criteris de ponderació i millorar la plataforma implementant noves funcionalitats.

Agraïments

Aquest treball ha estat patrocinat per "Ministerio de Sanidad y Consumo, Instituto de Salud Carlos III, Fondo de Investigación Sanitaria, nº 00/0033-02". Els resultats d'aquest treball han estat obtinguts utilitzant l'equip subvencionat per la Direcció de Recerca de la Generalitat de Catalunya (D.O.G.C 30/12/1997). Volem agrair a Enginyeria i Arquitectura La Salle el seu suport al nostre grup de Recerca en Sistemes Intel·ligents.

Referències

[1] A. Aamodt and E. Plaza. Case-Based Reasoning: Foundations Issues, Methodological Variations, and System approaches. *In AI Communications*, volume 7, pages 39-59, 1994.

[2] D. Aha. Feature weighting for lazy learning algorithms. *Technical Report AIC-98-003*, Navy

Center for Applied Rearch in AI, Washington, D.C. ,1998.

[3] D. Aha and P. Harrison. Case-Based Sonogram Classification. *Technical Report AIC-93-041*, Navy Center for Applied Research in AI, Washington, D.C. , October 1994.

[4] K-D. Althoff et al. Induction and Case-Based Reasoning for Classification Tasks. In H.H. Bock, W. Lenski, and M.M. Richter editors, *Information Systems and Data Analysis, Prospects-Foundations-Applications, Proc. 17th Annual Conference of the GFKL*, pages 3-16. University of Kaiserslautern, Springer-Verlag, Berlin-Heidelberg, 1994.

[5] P. Domingos. Context-sensitive feature selection for lazy learners. *Artificial Intelligence Review*, Volume 11, pages 227-253, 1997.

[6] J.M.Garrell, E. Golobardes, E.Bernadó and X. Llorà. Automatic diagnosis with Genetic Algorithms and Case-Based Reasoning. *In Artificial Intelligence Engineering*, volume 13, pages 357-362, October 1999, Elsevier Science Ltd. ISSN 0945-1810.

[7] J.M. Garrell, E. Golobardes, E. Bernadó and X. Llorà. Automatic Classification of Mamary Biopsy Images with Machine Learning techniques. *In proceedings of Engineering of Intelligent Systems* (EIS'98), 1998.

[8] M. Gawri's and J. Sienkiewicz. Rough Set Library user's Manual. Technical Report 00-665, *Institute of Computer Science*, Warsaw University of Technology, 1993.

[9] E. Golobardes and J.M. Garrell. Avaluació d'un sistema classificador basat en casos per a la diagnosi de biòpsies de teixit de glàndules mamàries. *In Proceedings de les Jornades d'Intel·ligència Artificial* (JIA'97), 1997.

[10] N. Howe and C. Cardie. Examining locally varying weights for neares neighbour algorithms. *In proceedings of the Second ICCBR Conference,* pages 456-466. Springer-Verlag, Providence, 1997.

[11] J. Kolodner. Case-Based Reasoning. *Morgan Kauffmann Publishers*, Inc., 1993.

[12] X. Llorà, E. Golobardes, and M. Salamó. Diagnosis of microcalcifications using Case-Based Reasoning and Genetic Algorithms. *In Proceedings of Engineering of Intelligent Systems* (EIS'2000), Volume 1, pages 254-263, 2000.

[13] J. Martí, J. Español, E. Golobardes, J. Freixenet, R. García, and M. Salamó. Classification of microcalcifications in digital mammograms using case-based reasoning. *In International Workshop on digital Mammography*, 2000.

[14] Z. Pawlak. Rough Sets. *In International Journal of Information and Computer Science*, Volume 11, 1982.

[15] Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data. *Kluwer Academic Publishers*, 1991.

[16] C. K. Riesbeck and R.C. Shank. Inside Case-Based Reasoning. *Lawrence Erlbaum Associates*, Hillsdale, NJ, US, 1989.

[17] M. Salamó, E. Golobardes, J.M. Garrell and J. Martí. Classificación de microcalcificaciones usando Razonamiento Basado en Casos. *In III Jornadas de Transferencia Tecnológica de Inteligencia Artificial* (TTIA'99), Murcia, Spain, November, 1999.

[18] M. Salamó, E. Golobardes, D. Vernet, and M. Nieto. Weighting methods for a Case-Based Classifier System. *In LEARNING'00*, Madrid, Spain, October 2000. IEEE.

[19] L. Shen, R.M. Rangayyan and J.L. Dessautels. Detection and classification of mammographic calcifications. *In State of the Art in Digital Mammographic Image Analysis*, pages 198-212, 1994.

[20] A. Skowron. Synthesis of Adaptive Decision Systems from Experimental Data. *In A. Aamodt and J. Komorowski, editors. Fith Scandinavian Conference on Artificial Intelligence*, volume 28, Trondheim, Norway, May 1995, IOS Press.

[21] I. Watson. Applying Case-Based Reasoning: Techniques for Enterprise Systems. *Morgan Kaufmann Publishers* Inc. 1997.