Generació de funcions de similitud mitjançant la Programació Genètica pel Raonament Basat en Casos

Elisabet Golobardes[†], Mireya Nieto[†], Maria Salamó[†], Joan Camps[†], Gemma Calzada[†], Joan Martí[‡] i David Vernet[†]

[†] Grup de Recerca en Sistemes Intel·ligents, Departament d'Informàtica, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Passeig Bonanova 8, 08022 - Barcelona.
[‡]Institut d'Informàtica i Aplicacions, Universitat de Girona, Avda. Lluís Santaló s/n, 17071 - Girona
{elisabet,mireyan,joanc,mariasal,gemmac,dave}@salleURL.edu; joanm@eia.udg.es

Resum

El raonament basat en casos recupera casos utilitzant una funció de similitud. La funció de similitud permet extraure els casos més similars de la memòria de casos al nou cas que volem resoldre. Per aquest motiu, és un dels punts centrals de tot el cicle del raonament basat en casos. Aquest article proposa l'ús de la programació genètica per tal de cercar funcions de similitud per al raonament basat en casos, ad hoc a un domini concret. L'avaluació d'aquesta proposta es realitza en el domini de diagnòstic de càncer de mama a partir de les microcalcificacions que es troben en una mamografia. Els resultats obtinguts per aquest domini es comparen amb diferents funcions de similitud utilitzades en el raonament basat en casos.

Paraules clau: raonament basat en casos, computació evolutiva, aprenentatge artificial, aplicacions generals de la IA en la medicina.

1 Introducció

Els sistemes de raonament basat en casos seleccionen casos emmagatzemats per a aconseguir solucionar nous casos. La selecció de casos es basa principalment en una funció de similitud. Malgrat tot, les funcions de similitud usades més conegudes en la literatura no recuperen correctament en alguns dominis. Això provoca la generació de funcions de similitud ad hoc al problema en concret que s'està tractant, normalment donades per l'expert del sistema. Aquest problema es presenta en més gran mesura, en dominis de planificació on l'adaptació dels casos és més necessària o en sistemes classificadors on la definició de les diferents classes està relacionada.

Les funcions de similitud es poden utilitzar en moltes àrees comuns o properes a la Intel·ligència Artificial. La primera àrea la trobem en el raonament analògic, en sistemes que es poden catalogar dins de la familia: $\{\mathbf{case}, \mathit{exemplar}, \mathit{instance}, \mathit{memory}\}$ **– based –** $\{\mathit{learning}, \mathbf{reasoning}\}$ [25, 2, 27]. Un altra àrea on s'utilitzen les funcions de similitud és en algunes xarxes neuronals, entre les que es situen les xarxes CounterPropagation [11], ART [7], els Mapes Autoorganitzatius [15]. També es poden trobar en altres camps com són: l'estadística, el reconeixement de patrons i la psicologia cognitiva. En qualsevol de les àrees d'aplicació, la funció de similitud pot fer esbiaixar el sistema d'aprenentatge.

Aquest article proposa un mètode automàtic de generació de funcions de similitud, adaptables al domini que s'està tractant, i revisades pel sistema de raonament basat en casos que les executarà. En concret, aquest sistema generador de funcions de similitud està basat en programació genètica (GP) i el sistema avaluador de les funcions de similitud generades és el raonament basat en casos (CBR).

L'article s'ha estructurat en els apartats descrits a continuació. En primer lloc, es realitza un resum del treball relacionat. La següent secció explica la unificació del sistema generador de funcions de similitud utilitzant un avaluador extern. La secció 5 mostra l'experimentació i l'anàlisi de resultats obtinguts per al nostre domini. Finalment es presenten les conclusions i línies futures.

2 Treball relacionat

Al llarg de la literatura podem trobar moltes funcions de similitud basades en una mesura de distància. Moltes d'aquestes funcions de similitud s'adapten molt bé en dominis on els casos estan representats com un conjunt d'atributs numèrics, però tenen problemes amb dominis definits amb atributs nominals.

Entre les funcions de similitud basades en distància trobem: la mètrica de Minkowsky [5], Mahalanobis [21], Camberra, Chebychev, Quadratic, Correlation, i Chi-quadrat [20], *Context-Similarity measure* [6], *Contrast Model* [29], les funcions basades en hyperrectangles [27, 8], *heterogenous distance functions* [30], entre d'altres.

Però entre totes les funcions de similitud que han aparegut, la més coneguda és la distància Euclidiana, la qual es defineix com:

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

On Cas_x i Cas_y són els dos casos que volem calcular la similitud; F is el nombre d'atributs que descriuen el cas; x_i, y_i representen el valor *i-èssim* per l'atribut per al Cas_x i Cas_y respectivament; i w_i és el pes de l'atribut *i-èssim*.

Una alternativa a aquesta funció és la distància de Manhattan, ja que requereix un temps computacional menor.

$$Sim(Cas\_x, Cas\_y) = \sum_{i=1}^{F} |x_i - y_i| \quad (2)$$

Les funcions Euclidiana i Manhattan són equivalents a la mètrica de Minkowsky (Bachelor, 1978) amb valors r=*2* i r=*1*, respectivament.

L'ús de GPs per a obtenir una funció que modelitzi el comportament d'un sistema a partir d'algunes dades està aplicat extensament a multitud de problemes (regressió simbòlica). Tot i així, la idea d'utilitzar aspectes del raonament basat en casos i la programació genètica conjuntament no està tan extensa. Cal destacar l'extracció d'atributs més rellevants amb programació genètica per tal de ser aplicats al CBR [4]. Aquesta extracció es realitza usant les *automatically defined functions* (ADFs) [16, 17] i es basa en els treballs previs realitzats per [24].

Altres treballs relacionats al raonament basat en casos i la computació evolutiva en aquest mateix sentit són els que usen els algorismes genètics com a extractors de les característiques més rellevants [28, 14]. Tots aquest treballs es basen en extreure els atributs més rellevants però no en buscar directament la funció de similitud *ad hoc* al problema, com es proposa en el present article.

3 Mètodes d'aprenentatge artificial

Prèviament a proposar el sistema híbrid, en aquesta secció presentarem breument els dos paradigmes d'aprenentatge artificial que hi intervenen: el *raonament basat en casos* i la *programació genètica.*

3.1 El raonament basat en casos

El raonament basat en casos (*Case-Based Reasoning*) integra l'aprenentatge artificial i la resolució de problemes. Aquesta metodologia utilitza una filosofia similar a la que usem els humans: intenta resoldre els nous casos d'un problema a partir de casos ja resolts anteriorment [25]. La resolució d'aquests nous casos aporta informació i coneixement per tal de resoldre altres casos en un futur. Clàssicament, aquest mètode ve caracteritzat per quatre fases [1]: una primera fase de *recuperació* (*Retrieval phase*) de casos similars al nou cas que ja hagin estat resolts anteriorment; una segona fase en què s'intenta *adaptar* (*Reuse phase*) la solució dels casos anteriors recuperats, per tal de resoldre el nou cas; una tercera fase en què es *revisa* (*Revise phase*) la solució proposada; i, finalment, una quarta fase en què s'*emmagatzema* (*Retain phase*) la informació rellevant obtinguda d'aquesta resolució.

En un sistema de raonament basat en casos, una de les fases que juga un dels papers més importants, per tal d'assolir un bon rendiment del sistema classificador, és la fase de recuperació i, més concretament, les funcions de similitud. En aquest sentit, sorgeix la idea d'incorporar la programació genètica dins del CBR. És a dir, es pretén usar la programació genètica per tal de cercar les funcions de similitud més adequades per a cada domini que es vulgui resoldre.

3.2 La programació genètica

La programació genètica (GP o *Genetic Programming*) [16] és una tècnica d'intel·ligència artificial que, de la mateixa manera que els algorismes

genètics (GA o *Genetic Algorithms*) [12], es basa en el principi de l'evolució natural d'en Darwin i les lleis de l'herència d'en Mendel, fonament de la computació evolutiva [12]. Existeix una població d'individus que evolucionen al llarg del temps gràcies a la recombinació sexual -o creuament- i a la mutació de la informació genètica. Cada individu s'adapta més o menys bé al medi, té diferent valor d'avaluació -o de *fitness*-, i la selecció natural provoca que els millors individus tendeixin a perdurar al llarg de les generacions. La informació genètica codifica les diferents solucions a un problema, i l'individu que millor s'adapta al medi és considerat la millor solució.

En la població d'individus es van succeint diferents generacions al llarg del temps. Cada generació experimenta una sèrie de fases per tal de convertir-se en la generació següent. La primera fase és l'*avaluació*, on es calcula quin és el grau d'adaptació de l'individu al medi. La segona fase és la *selecció*, on s'escullen els individus més ben adaptats. Aquests passen a la fase de *creuament*, en la qual es produeix una recombinació del material genètic de dos individus per a formar-ne dos de nous. I, finalment, la *mutació* fa que hi hagi petits canvis aleatoris en el material genètic.

La diferència entre GA i GP és la forma en la qual es representa la informació genètica dels individus d'una població. En GP els individus són considerats programes i es representen en forma d'arbre. En conseqüència, els operadors genètics de *creuament* i de *mutació* s'adapten a aquesta representació. En Koza [16] va proposar aquest tipus de representació i els seus operadors genètics al definir en què consistia la programació genètica.

Una de les utilitats de la GP -segons Koza- és el descobriment de les lleis empíriques, o bé, el descobriment d'equacions. Això és possible amb la utilització de la regressió simbòlica com a estratègia fonamental. Amb la regressió, a partir d'un conjunt de dades o de mostres recollides empíricament sobre un sistema, es pot trobar la llei o equació que les governa utilitzant la GP. Només cal que els individus representin equacions, i el *fitness* de cadascun d'ells dependrà de com de bé s'ajusti el càlcul que se n'obté respecte el valor que s'esperava.

En aquest treball també es pretén trobar una funció, però la forma de trobar el *fitness* no serà a partir de les mateixes dades sinó del resultat d'aplicar-les al raonament basat en casos, usant com a *funció de similitud* l'equació que s'estigui avaluant en aquell moment.

4 Cerca de funcions de similitud usant GP

En aquesta secció proposem el sistema híbrid per tal de cercar funcions de similitud, *ad hoc* a un domini, per a un sistema raonador basat en casos usant la programació genètica.

4.1 Descripció del sistema híbrid

El sistema híbrid entre el raonament basat en casos i la programació genètica que proposem, té com a objectiu buscar les millors funcions de similitud per a un sistema raonador basat en casos *ad hoc* a un domini. Aquest sistema cerca les funcions en qüestió de forma automàtica i adaptativa. *Automàtica* perquè no serà un investigador qui proposi una nova funció, sinó que s'utilitzarà un mètode de cerca automàtic: la programació genètica. I *adaptativa* perquè, amb aquest sistema, per cada problema nou plantejat, es pot buscar quina és la funció que millors resultats dóna.

Aquest sistema híbrid usa les dues tècniques, la programació genètica i el raonament basat en casos de forma conjunta, i es pot veure tant des del punt de vista que la programació genètica és una utilitat pel raonament basat en casos, com des del punt de vista que el CBR és una part de la GP. Analitzem els dos enfocaments.

4.1.1 El sistema híbrid des del punt de vista del CBR

Des del punt de vista del CBR la GP és una eina que actua com una *caixa negra*, proposant funcions de similitud. Així, dins de l'etapa de recuperació del cicle del CBR, quan cal utilitzar la funció de similitud es consulta al GP quina és la funció que s'ha d'utilitzar. Vegeu la figura 1.

Concretament, la part del sistema híbrid corresponent a la programació genètica lliura la funció de similitud. Aquesta funció s'analitza amb un intèrpret de funcions i es carrega a memòria. A partir d'aquest moment, la part corresponent al CBR funciona normalment realitzant el seu cicle sobre tots els casos que constitueixen el problema a resoldre, de forma que en tot moment la funció de similitud utilitzada és la que el GP li ha lliurat inicialment.

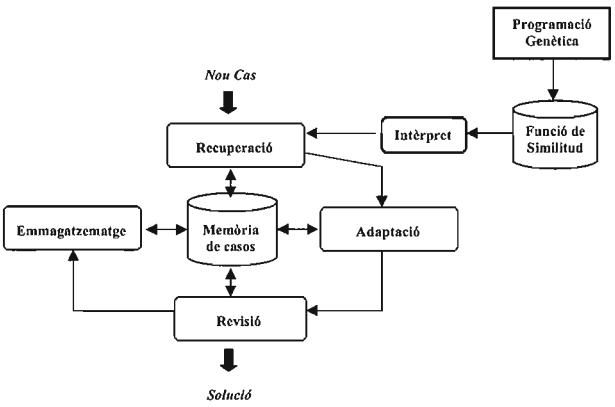

Figura 1: Cicle del raonament basat en casos en el sistema híbrid.

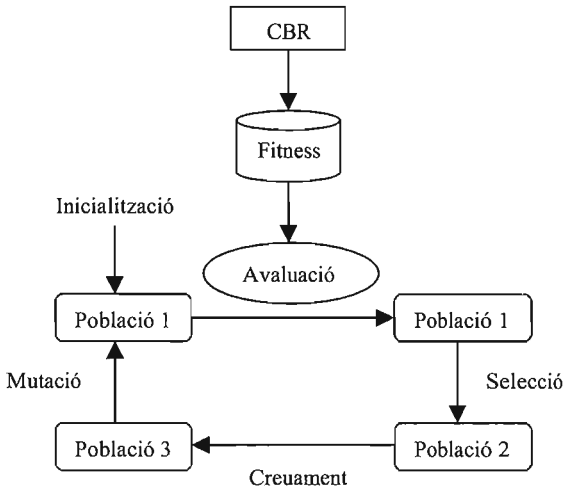

Figura 2: Cicle de la programació genètica en el sistema GP-CBR.

4.1.2 El sistema híbrid des del punt de vista del GP

Un altre punt de vista és considerar que el CBR és una eina que s'utilitza en una fase del GP. Durant la fase d'avaluació del GP, es realitza una crida a un altre sistema, la part corresponent al CBR, per tal de calcular quin és el *fitness* de cada individu. Així, la fase d'avaluació és un sistema extern al GP i que està basat en un CBR que actua com una *caixa negra*, i que retorna en un fitxer l'estadística del percentatge d'encerts que s'ha aconseguit per cada individu -és a dir, funció de similitud proposada-. Vegeu la figura 2.

Resumint, per cada generació, quan el GP arriba a la fase d'avaluació, es comuniquen al CBR quines són les funcions -la població- a avaluar. L'avaluació de cada individu consisteix en l'execució de tot el cicle de CBR per cadascuna de les funcions sobre un problema concret, i el resultat obtingut es retorna al GP. Quan es posseeix el *fitness* de totes les funcions d'aquella població, es continua amb el cicle del GP.

4.2 Les noves funcions de similitud

El sistema híbrid proposat cerca, mitjançant la GP, noves funcions de similitud. Així doncs, cada individu de la població genètica és una funció de similitud que es posa a prova com a possible candidata per a una bona solució. Serà una bona solució aquella funció que faci una bona classificació al ser utilitzada en la part corresponent al CBR. A continuació descrivim com són les funcions de similitud candidates.

Els individus de la GP es representen en forma d'arbres. Aquests arbres estan formats per nodes anomenats *funcions* i nodes anomenats *terminals*. Per tal de no confondre la funció de similitud amb les funcions dels nodes dels arbres, anomenarem a aquests últims *nodes-funció*. Els nodes-funció i els terminals poden prendre un valor possible dins del *conjunt de funcions* (o millor dit *conjunt de nodes-funció*) i del *conjunt de terminals* respectivament. Per escollir el conjunt de nodes-funció i de terminals s'han de tenir en compte dues condicions: el *tancament* i la *suficiència*. Aquestes condicions exigeixen que els conjunts escollits siguin compatibles entre si en quant al valor i al tipus de dades que retornen, i que sigui suficient per tal de poder construir la solució al problema.

Un possible conjunt de terminals T i de nodes-funció $N-F$ que compleix aquests requeriments per a crear funcions de similitud -*ad hoc* al problema de diagnòstic a partir de les microcalcificacions de les mamografies (vegeu la secció 5.1)- és:

$T = \{X1, X2, .., X21, Y1, Y2, .., Y21, R\}$

$N\text{-}F = \{+, -, \times, /, log, absol, arrel2, arrel3, exp2, exp3, exp4, suma21\}$

Tal que el significat dels símbols és:

- $X1, .., X21$: els atributs que descriuen les microcalcificacions d'un cas conegut
- $Y1, .., Y21$: els atributs que descriuen les microcalcificacions d'un cas nou

- K: nombres reals entre 0 i 1
- $+, -, \times, /$: funcions aritmètiques suma, resta, producte i divisió
- *log*: logaritme
- *absol*: valor absolut
- $arrel2, arrel3$: arrel quadrada i arrel cúbica, respectivament
- $exp2, exp3, exp4$: elevar al quadrat, al cub o a la quarta potència, respectivament
- $suma21$: sumatori de 21 membres

El criteri que s'ha seguit per escollir aquests nodes-funció ha estat que el GP fos capaç de generar les funcions de similitud ja conegudes, les descrites en la secció 2. La intenció és que els resultats preliminars corroborin el fet de què el GP és capaç de generar -com a mínim- les funcions de similitud de propòsit general ja conegudes, i les seleccioni com a bones candidates, tot i que pot ser que en generi alguna totalment diferent a qualsevol de coneguda.

També es podria haver escollit uns nodes-funció més generals, per exemple, que hi hagués un node-funció arrel amb la potència com a paràmetre, però en GP és més important restringir l'espai de cerca al màxim per tal de reduir el temps computacional, sempre tenint en compte que es satisfaci la suficiència. Una altra alternativa seria definir funcions més concretes, per exemple, un node-funció que fes el sumatori del valor absolut de la diferència dels atributs i que només deixés lliure el valor dels pesos. L'inconvenient és que es restringeix massa l'espai de cerca, i estem dificultant el descobriment de les funcions de similitud que siguin completament diferents a les conegudes.

Un exemple de com seria en GP l'individu que representaria la funció de similitud de la mètrica de Minkowski s'il·lustra a la figura 3.

5 Experimentació

En aquesta secció volem analitzar -de manera preliminar- els resultats que s'obtenen amb el sistema híbrid. En primer lloc, descrivim el problema i, en segon lloc, analitzem els resultats obtinguts.

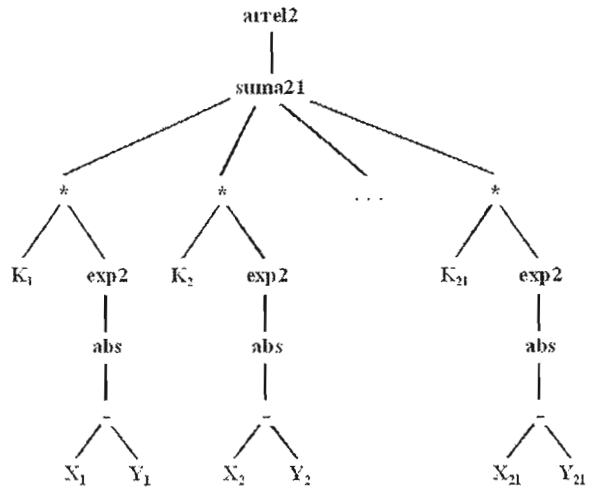

Figura 3: Possible funció que s'esperaria que generés el sistema híbrid.

5.1 Definició del problema

El nostre problema consisteix en la detecció de càncer de mama a partir de les microcalcificacions trobades en una mamografia. La incidència del càncer de mama vària depenent del país, però les últimes estadístiques mostren que aproximadament 720.000 casos nous es diagnostiquen en tot el món, d'aquests casos un 20% són malignes. La tècnica més utilitzada per detectar el càncer de mama són les mamografies degut a la seva alta sensibilitat en trobar tumors petits. Normalment les microcalcificacions i el creixement de massa són el primer signe de càncer de mama.

Les microcalcificacions normalment apareixen com petits punts brillants situats de forma arbitrària en la mama i amb formes també molt diverses. El que s'ha fet és extraure les característiques més rellevants a partir dels descriptors de visibilitat mitjançant tècniques de processat d'imatges. Aquest procés l'ha realitzat la Universitat de Girona [18] amb col·laboració amb l'Hospital de Girona "Doctor Josep Trueta".

Les figures 4 i 5 corresponen a la imatge original i segmentada respectivament. La imatge original és segmentada per tal d'aconseguir una millor visibilitat de les microcalcificacions. Aquesta mamografia segmentada serà la que s'utilitzarà per a treure els descriptors de forma de les microcalcificacions.

La Universitat de Girona ens ha cedit les dades numèriques de les mamografies digitalitzades. Les dades representen les característiques basades en la

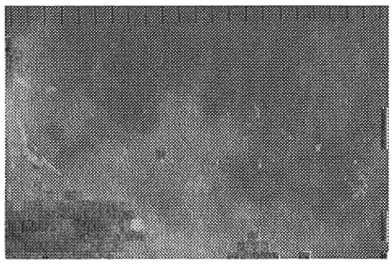

Figura 4: Imatge mamografia original.

Figura 5: Imatge mamografia segmentada.

forma que es pot apreciar de cadascuna de les microcalcificacions trobades en una mamografia. A partir d'aquestes dades es vol detectar si una persona pateix o no càncer de mama [10].

Disposem de dos conjunts de dades (*training* i *test*) on cada conjunt té un conjunt de mamografies representades per un nombre indeterminat de microcalcificacions, de les quals disposem de 21 característiques a avaluar. El conjunt d'entrenament està format per 216 casos i el de test per 70.

5.2 Resultats preliminars

Per tal d'analitzar el rendiment de les funcions proposades pel GP, tot seguit analitzem dos tipus d'experiments. En primer lloc, mantenim el conjunt d'entrenament i de test proposat pels experts humans -metges- [18]. Així doncs, comparem les funcions proposades respecte dels resultats obtinguts pels experts humans (`H-E`), respecte del model estadístic (`SM`) que es va proposar en una anàlisi prèvia [18], respecte de les funcions clàssiques del raonament basat en casos (`Hamming`, `Euclidiana`, `Cúbica` i `Clark`) [19, 10], i respecte de les funcions proposades pel sistema híbrid. A tall d'exemple analitzem dues propostes diferents: la `Funció1-GP` (vegeu l'equació 3) i la `Funció2-GP` (vegeu l'equació 4). Podem observar que la `Funció1-GP` és una espècie de Hamming on els pesos a vegades són valors reals i d'altres són relacions entre atributs. Fins i tot, per la diferència $X_{19} - Y_{19}$ el pes és de nou un sumatori. Mentre que la `Funció2` sembla "excessivament"senzilla. En aquest cas, el que fa el GP és buscar una funció que classifiqui principalment els casos majoritàriament representats en el conjunt d'entrenament, tendència que s'ha d'evitar. Vegeu la taula 1.

Realitzem aquesta comparació basant-nos en el percentatge de casos *sense classificar*, en la *sensitivitat* i en l'*especificitat*, i en el percentatge d'*encerts*. Entenem que la *sensitivitat* és la proporció dels *vertaders positius* (casos malignes), i l'*especificitat* és la relació dels *vertaders negatius*.

Podem observar que, tant les funcions proposades pel sistema híbrid, com les funcions basades en la distància del CBR obtenen resultats molt similars. Malgrat que el percentatge d'encerts és superior a l'obtingut pels metges, o bé, pel model estadístic, la fiabilitat (sensitivitat i especificitat) és molt menor. Així doncs, s'ha de seguir treballant en *descobrir* d'una banda millors funcions i, d'altra banda, en trobar *llindars* de fiabilitat dins del CBR. És a dir, és preferible no diagnosticar -classificar- a realitzar un diagnòstic incorrecte.

En segon lloc, analitzem els resultats obtinguts respecte d'altres tècniques conegudes d'aprenentatge artificial, usant la *stratified ten-fold cross-validation*. Aquests algorismes s'han extret de l'entorn *Weka* [31], disponible a l'adreça http://www.cs.waikato.ac.nz/ml/weka. Els algorismes escollits són: els *Instance-based learning* `IB1` i `IB3` [3]; el model estadístic *Naive Bayes* (`NB`) [13]; el basat en arbres d'inducció `C4.5 revision 8` [23]; el basat en aprenentatge basat en regles `PART` [9]; i el basat en màquines de suport vectorial `SMO` [22]. Tots ells s'han usat amb la configuració que ofereix per defecte el Weka. En aquesta segona anàlisi dels resultats, no oferim el percentatge de no classificats, ja que les versions que es presenten classifiquen tots els casos, mentre que la *sensitivitat* i l'*especificitat* són la mitjana de les execucions dels *ten-folds* (vegeu la taula 2).

Variant	No clas.	Sens.	Espec.	Encerts
H-E	38.57	70.59	92.59	52.86
SM	52.86	81.82	90.48	40.00
Hamming	0.00	80.95	69.39	72.86
Euclidiana	0.00	66.67	75.51	72.86
Cúbica	0.00	66.67	77.55	74.29
Clark	0.00	66.66	81.63	77.14
Funció1-GP	0.00	76.19	69.38	71.43
Funció2-GP	0.00	4.76	97.95	70.00

Taula 1: Resultats obtinguts usant el conjunt d'entrenament i de test originals.

$$
\begin{aligned}
Sim(Cas\_X, Cas\_Y) = {} & \sqrt[3]{X_{15}}(X_1 - Y_1) + 7.255(X_2 - Y_2) + \\
& (X_{19} - Y_{12})(X_3 - Y_3) + \sqrt[3]{Y_{10}}(X_4 - Y_4) + \\
& {Y_{18}}^4(X_5 - Y_5) + 0.772(X_6 - Y_6) + \\
& (8.163 - X_{13})(X_7 - Y_7) + {Y_{18}}^4(X_8 - Y_8) + \\
& \sqrt[3]{X_{19}}(X_9 - Y_9) + 0.758(X_{10} - Y_{10}) + \\
& {Y_{16}}^4(X_{11} - Y_{11}) + (1.440 + X_4)(X_{12} - Y_{12}) + \\
& {Y_{11}}^2(X_{13} - Y_{13}) + (0.132 - Y_4)(X_{14} - Y_{14}) + \\
& (Y_{18} + Y_{16})(X_{15} - Y_{15}) + 1.209(X_{16} - Y_{16}) + \\
& 13.368(X_{17} - Y_{17}) + \frac{X_7}{X_{20}}(X_{18} - Y_{18}) + \\
& (Y_3(X_1 - Y_1) - 6.330(X_2 - Y_2) + X_{12}(X_3 - Y_3) + \\
& Y_4(X_4 - Y_4) + Y_{15}(X_5 - Y_5) + 7.013(X_6 - Y_6) + \\
& X_9(X_7 - Y_7) + Y_5(X_8 - Y_8) + X_2(X_9 - Y_9) + \\
& Y_{20}(X_{10} - Y_{10}) + X_{20}(X_{11} - Y_{11}) + Y_{20}(X_{12} - Y_{12}) + \\
& X_{15}(X_{13} - Y_{13}) + Y_6(X_{14} - Y_{14}) + \\
& 8.341(X_{15} - Y_{15}) + 5.999(X_{16} - Y_{16}) + \\
& 4.293(X_{17} - Y_{17}) + 7.055(X_{18} - Y_{18}) + \\
& 4.871(X_{19} - Y_{19}) + X_7(X_{20} - Y_{20}) - \\
& 3.534(X_{21} - Y_{21}))(X_{19} - Y_{19}) + \\
& \sqrt[3]{Y_{21}}(X_{20} - Y_{20}) + (Y_6 + 3.558)(X_{21} - Y_{21})
\end{aligned}
\tag{3}
$$

$$
Sim(Cas\_X, Cas\_Y) = Y_{18} + X_9 - Y_{15} - (Y_5 + X_3)^3 \tag{4}
$$

Donada la taula 2 podem observar resultats força similars des de diferents algorismes de l'aprenentatge artificial, pel que fa a la seva capacitat de diagnòstic davant d'aquest problema. Així doncs, es necessita realitzar el mateix treball immediat que se'n dedueix de la taula 1.

Val a dir, que la cerca de funcions de similitud de funcions mitjançant el sistema híbrid proposat, té un alt cost computacional (de l'ordre de 25 a 30 hores i, en alguns, casos 3-4 dies). Tot i que s'ha de millorar el rendiment del sistema generador de funcions, val a dir, que una vegada s'ha trobar una

Variant	Sens.	Espec.	Encerts	Std
Hamming	56.84	66.94	62.50	14.47
Euclidiana	56.84	69.42	63.89	12.43
Cúbica	60.00	68.60	64.81	9.62
Clark	57.10	62.88	60.40	12.73
Funció1-GP	55.79	67.77	62.45	12.74
Funció2-GP	9.47	91.73	55.48	3.50
IB1	56.84	67.77	62.96	12.42
IB3	60.00	69.42	65.28	6.29
C4.5 (r8)	70.53	60.33	64.81	6.36
NB	60.00	68.60	64.81	7.66
PART	80.00	47.93	62.04	4.17
SVM-SMO	52.63	78.51	67.13	7.37

Taula 2: Resultats obtinguts usant *10-fold cross-validation.*

bona funció, el que compte és el cost del sistema basat en casos. En la configuració usada per aquest treball, la nostra plataforma [26] té un cost de CPU -en mitjana- de 649 milisegons per resoldre un cas nou.

6 Conclusions i treball futur

Aquest article té com a principal objectiu presentar una aproximació a un generador de funcions de similitud, *ad hoc* al domini que s'està tractant, per a un sistema de raonament basat en casos. Aquesta proposta s'ha avaluat dins del domini de diagnòstic de càncer de mama a partir de les microcalcificacions trobades en una mamografia. Els primers resultats obtinguts a partir de funcions de similitud, generades pel sistema híbrid, han demostrat la viabilitat del generador de funcions per a problemes on el domini especificat és difícil de tractar amb les funcions més usades dins del CBR. L'anàlisi d'aquests resultats amb altres funcions de similitud típiques del CBR han demostrat, a la seva vegada, la funcionalitat del sistema.

Com a línies de treball futures es proposen: (1) la millora del rendiment en temps del sistema generador de funcions; (2) la modificació dels operadors del generador per aconseguir funcions de similitud més acurades; (3) l'estudi de la generalització de les funcions de similitud obtingudes; i finalment (4) l'estudi d'aquest sistema híbrid per d'altres dominis.

Agraïments

Voldríem agrair al *Ministerio de Sanidad y Consumo, Instituto de Salud Carlos III, Fondo de*

Investigación Sanitaria pel finançament del projecte FIS 00/0033. Els resultats d'aquest treball s'han obtingut usant l'equip co-finançat per la *Direcció General de Recerca de la Generalitat de Catalunya (D.O.G.C 30/12/1997)*. Finalment, voldríem agrair a Enginyeria i Arquitectura La Salle, de la Universitat Ramon Llull, pel 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 and P. Harrison. Case-Based Sonogram Classification. Technical Report AIC-93-041, NRL/FR/5510-94-9707, Navy Center for Applied Research in AI, Washington, D.C., October 1994.

[3] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning, Vol. 6*, pages 37–66, 1991.

[4] M. Ahluwalia and L. Bull. Coevolving functions in genetic programming: Classification using k-nearest-neighbour. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 1999.

[5] B. Bachelor. Pattern recognition: Ideas in practice, 1978.

[6] Yoram Biberman. A context similarity measure. In *European Conference on Machine Learning*, pages 49–63, 1994.

[7] Gail A. Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37(1):54–115, 1987.

[8] P. Domingos. Context-sensitive feature selection for lazy learners. In *AI Review*, volume 11, pages 227–253, 1997.

[9] E. Frank and I. H. Witten. Generating Accurate Rule Sets Without Global Optimization. In *Machine Learning: Proceedings of the Fitfteenth International Conference*, pages 152–160. Kluber, 1986.

[10] E. Golobardes, X. Llorà, M. Salamó, and J. Martí. Computer Aided Diagnosis with Case-Based Reasoning and Genetic Algorithms. *Journal of Knowlegde Based Systems, Elsevier Science Ltd.*, page In Press, 2001.

[11] R. Hecht Nielsen. Counter propagation network. *Applied Optics, Vol 26, No23, Dec,1987*, 1987.

[12] J. H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* MIT Press/ Bradford Books edition, 1975.

[13] G. H. John and P. Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in AI*, pages 338–345. Morgan Kaufman Publishers, 1995.

[14] J. D. Kelly and L. Davis. Hybridizing the genetic algorithms and the k nearest neighbors. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, 1991.

[15] T. Kohonen. The Self-Organizing Map. In *New Concepts in Computer Science: Proc. Symp. in Honour of Jean-Claude Simon*, pages 181–190, Paris, France, 1990. AFCET.

[16] J.R. Koza. *Genetic Programming. On the Programming of Computers by Means of Natural Selection.* MIT Press, 1992.

[17] J.R. Koza. *Genetic Programming II. Automatic Discovery of Reusable Programs.* MIT Press, 1994.

[18] J. Martí, X. Cufí, J. Regincós, and et al. Shape-based feature selection for microcalcification evaluation. In *Imaging Conference on Image Processing, 3338:1215-1224*, 1998.

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

[20] R. Michalski, R. Stepp, and E. Diday. A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts, 1981.

[21] M. Nadler. Pattern recognition engineering, 1993.

[22] J. C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1998.

[23] R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[24] M.L. Raymer, W.F. Punch, E.D. Goodman, and L.A. Kuhn. Genetic programming for improved data mining: An application to the biochemistry of protein interactions. In *Genetic Programming 1996: Proceedings of the First Annual Conference*, pages 375–380. MIT Press, 1996.

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

[26] M. Salamó and E. Golobardes. BASTIAN: Incorporating the Rough Sets theory into a Case-Based Classifier System. In *III Congrés Català d'Intel·ligència Artificial (CCIA'00)*, 2000.

[27] S. Salzberg. A nearest hyperrectangle learning method. *Machine Learning*, 6:277–309, 1991.

[28] W. Siedlecki and J. Sklansky. On automatic feature selection. In *Proceedings of the International Journal of Pattern Recognition and Artificial Intelligence*, 1988.

[29] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[30] D.R. Wilson and T.R. Martinez. Improved heterogeneus distance functions. *Journal of Artificial Intelligence Research (JAIR)*, 6(1):1–34, 1997.

[31] I. H. Witten and E. Frank. *DataMining: practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann Publishers, 2000.