Facultade de Informática
UNIVERSIDADE DA CORUÑA

TRABALLO FIN DE GRAO
GRAO EN ENXEÑARÍA INFORMÁTICA
MENCIÓN EN COMPUTACIÓN

Menos é máis: explorando o impacto da selección de características.

Estudante: Esteban Pichel Bolón
Dirección: Verónica Bolón Canedo
Laura Morán Fernández

A Coruña, xuño de 2020.

A ti, que remaches comigo cando o vento sopraba en contra.

Agradecementos

Grazas á miña familia, en especial a meus pais, por todo o apoio que me bridanron durante todo este camiño. Ás miñas titoras, Vero e Laura, polo bo trato e por toda a adicación mostrada na realización deste traballo. Aos meus compañeiros do grupo LIDIA e do grao, polos bos momentos vividos e a gran axuda que sempre me ofreceron. A Alberto, por todo o que puiden aprender del e con el. E aos meus amigos, por vivir todo esto comigo.

Resumo

Neste proxecto estúdanse diferentes técnicas de selección de características coa intención de determinar cales dos múltiples métodos existentes na literatura son máis axeitados para un tipo de problema en concreto, e determinar se algún deles é descartable por obter peores resultados que realizar unha selección aleatoria á hora de reducir a dimensionalidade dos problemas. Para a súa realización farase uso dun extenso número de conxuntos de datos que nos permitan traballar sobre unha gran variedade dos problemas existentes no mundo real cos que nos atopamos neste ámbito, reducindo a súa dimensionalidade e levando a cabo a clasificación correspondente para obter os resultados que, xunto cos test estadísticos, nos permitirán sacar conclusións sobre as cuestións plantexadas.

Abstract

In this project we study different feature selection techniques with the aim of determining which of the multiple methods in the literature are the best suited for a particular type of problem, and determining whether any of them are disposable because of obtaining worse results than a random selection to reduce the dimensionality of the problems. To accomplish this objective we will use an extensive number of data sets that allow us to work on a wide variety of problems from the real world that need to be dealt with in this field. We will reduce the dimensionality of the data sets and carry out the corresponding classification process to obtain the results that, along with statistical tests, will allow us to draw conclusions about the issues raised.

Palabras chave:

- Selección de características
- Filtros
- Redución da dimensionalidade
- Clasificación
- Alta dimensión
- Datasets microarray

Keywords:

- Feature selection
- Filters
- Dimensionality reduction
- Classification
- High dimensionality
- Microarray datasets

Índice Xeral

Índice de Figuras

Índice de Táboas

Capítulo 1

Introdución

Na actualidade a gran velocidade coa que se están a xerar datos e a importancia que ten a análise dos mesmos en diferentes ámbitos e aplicacións fan que xurda a necesidade de empregar técnicas de aprendizaxe automática e plataformas de Big Data para poder abordalos, e obter desta forma información e coñecemento de calidade.

A clasificación é unha tarefa que posúe gran importancia dentro da análise de datos, o recoñecemento de patróns e a aprendizaxe máquina. Existen centos de algoritmos de clasificación, pero a maioría deles presentan certa degradación no rendemento cando se enfrontan a moitas características irrelevantes e/ou redundantes, incluso cando estas son relevantes. É por esta razón pola que cobran importancia os métodos de preprocesado dos datos, neste caso a chamada selección de características, cuxo obxectivo é determinar o mellor subconxunto de características que describe precisamente un problema dado cunha degradación mínima do rendemento.

Os métodos de selección de características poden clasificarse segundo a súa relación co algoritmo de indución, tendo (i) filtros, que son independentes do algoritmo de indución e establecen a importancia das características en base a métricas como a información mutua ou estatísticos como Chi^2; (ii) métodos envolventes, que usan a precisión do algoritmo de indución para determinar a importancia de subconxuntos de características; e (iii) métodos embebidos, que realizan a selección durante o proceso de adestramento do algoritmo de indución. Asimesmo, os métodos de selección de características tamén se poden distinguir en univariados (cando estudan a relevancia dunha característica con respecto á clase predictiva); e multivariados (cando estudian as interaccións entre subconxuntos de características).

A selección de características é unha técnica de preprocesado amplamente utilizada polos analistas de datos, demostrando excelentes resultados nunha multitude de aplicacións. Son moitos e moi variados os distintos métodos de selección de características que nos atopamos na literatura especializada e non deixan de aparecer novos métodos cada ano; e é aquí onde nos xurde a cuestión de comprobar se realmente necesitamos tantos métodos de selección de

características e cales deles son certamente útiles.

Debido a isto, semella interesante realizar unha análise sobre os métodos de selección de características máis populares, comparalos entre sí e coa selección aleatoria (como liña de base), para determinar en base a tests estatísticos cales son os métodos máis eficaces e efectivos e descartar desta forma aqueles que sexan superados por unha reducción da dimensionalidade de forma aleatoria.

1.1 Obxectivos

O obxectivo entorno ao cal se desenvolve este traballo consiste en realizar un estudo exhaustivo dos distintos métodos de selección de características que existen na literatura, empregando unha gran variedade de conxuntos de datos que posúan gran parte dos problemas habituais cos que nos podemos atopar neste campo da investigación, e poder, deste xeito, dar unha recomendación de cales destes métodos son os que mellores resultados aportan, e polo tanto os recomendados para seren empregados dado un problema concreto. Como liña base tomaremos o método de selección aleatoria, sobre o cal se compararán o resto dos métodos, de xeito que permita evitar o uso daqueles métodos que non demostren ser mellores que facer unha redución da dimensionalidade do problema de forma aleatoria.

Por outra banda, empregaremos diferentes clasificadores cos cales obteremos os resultados que nos permitan avaliar os experimentos propostos. Deste xeito, poderemos comprobar a influenza da selección no rendemento dos distintos algoritmos de clasificación que se poidan empregar a posteriori, de xeito que se reduza a importancia á hora de seleccionar un ou outro clasificador.

1.2 Estudo de viabilidade e impacto

Previa realización dun proxecto, sexa do tipo que sexa, cómpre realizar un estudo da súa viabilidade e impacto, é dicir, deben realizarse os estudos e comprobacións pertinentes para asegurar que o proxecto é viable en termos de duración e custo. Para a súa realización, é necesario definir a totalidade de tarefas que se plantexan levar a cabo, o tempo que implicaría a correspondente execución, estimar os recursos dispoñibles e os custos asociados.

Nesta ocasión, o proxecto consta das seguintes fases que serán expostas a continuación. Non obstante, optouse por seguir unha metodoloxía áxil de refinamento progresivo, adaptándose aos problemas e retrasos que poidan xurdir.

- **Fase 1: Obtención da documentación.** Trátase de analizar e comprender cal é o ámbito do coñecemento abarcado polo problema en cuestión, de forma que realiza a

busca de toda aquela información que poida ser útil e sobre a que se basea o traballo, as tecnoloxías base a empregar e a análise doutras aproximacións de temática similar.

- **Fase 2: Análise de técnicas de selección de características.** Relizarase un estudo dos diferentes métodos existentes na literatura de forma que logremos identificar cales son os que máis nos interesan e mellor se adaptan ao problema que estamos a tratar.

- **Fase 3: Estudo dos algoritmos de clasificación.** Nesta fase buscarase información sobre os distintos clasificadores que se empregan en problemáticas semellantes, escollendo aqueles que consideremos que posúen as características que nos interesan e que se poidan usar na metodoloxía proposta así como de tests estatísticos para avaliar os resultados obtidos.

- **Fase 4: Estudo dos conxuntos de datos a empregar.** Accederase ao repositorio UCI[1] do cal se seleccionarán aqueles conxuntos de datos que, en función das súas características e dos métodos de seleción máis axeitados a empregar, cumplan as condicións que se buscan para o problema. Tamén se accederá a repositorios de conxuntos de datos de tipo microarray, para ter maior variabilidade nos problemas a tratar.

- **Fase 5: Deseño e implementación da metodoloxía.** Determinaranse os pasos a seguir de forma que permita avaliar o avance do proceso en comparación cunha liña base.

- **Fase 6: Análise de resultados e conclusións.** Unha vez realizados todos os experimentos desexados, tomaranse os resultados obtidos en dita experimentación para realizaren a correspondente análise e determinar cales foron as conclusións ás que se chegou en función dos memos.

- **Fase 7: Elaboración da memoria do proxecto.** Crearase un documento no que se plasme todo o proceso levado a cabo durante o proxecto.

- **Fase 8: Propoñer liñas de traballo futuro.** Unha vez rematadas todas as fases de implementación do proxecto, plantexaranse liñas futuras sobre as que se poida continuar partindo da investigación levada a cabo neste traballo.

Destas fases extráense as tarefas que se inclúen na planificación do proxecto e que se mostran no diagrama de Gantt da Figura 1.1. A elaboración da memoria do proxecto pode realizarse de forma simultánea con outras tarefas, co cal o recurso que lles for asignado non traballará nelas con adicación total, senón que repartirá o seu horario para a realización de ambas. O tempo necesario adicado, de forma parcial, para levar a cabo a realización de todas

[1] https://archive.ics.uci.edu/ml/index.php

as tarefas do proxecto foi de 212 días, comezando o 3 de decembro de 2019 e finalizando o 25 de xuño de 2020, supoñendo un total de 850 horas de traballo.

Figura 1.1: Diagrama de Gantt do proxecto.

Os recursos humanos necesarios para realizar o proxecto consistiron nun analista programador e dous directores. O salario bruto anual dun analista programador segundo o convenio está fixado a 24.157,27 euros a tempo completo, polo que o custo por hora é de 12,60 euros. Na Táboa 1.1 amósanse os recursos empregados no proxecto, ademais dos custos asociados a cada un deles. Pódese dicir que o proxecto é viable economicamente xa que o custo é asumible.

Tipo de recurso	**Nome**	**Custo**
Software	SO Microsoft Windows 10	118 €
	Matlab R2019a	800 €
	Weka	0 €
	Overleaf	0 €
Hárdware	PC	700 €
Recursos humanos	Enxeñeiro informático	10.710 €
Total		12.328 €

Táboa 1.1: Resumo dos custos do proxecto

1.3 Estrutura da memoria

Esta sección presenta un breve resumo dos capítulos nos que se divide a memoria deste proxecto:

- **Capitulo 1: Introdución.** Neste capítulo explícase a motivación do proxecto, así como tamén se delimita o seu alcance e obxectivos e se expón a metodoloxía empregada para levalo a cabo.
- **Capitulo 2: Materiais e métodos.** Este capítulo comenta e describe os materiais empregados neste proxecto fin de carreira, como as ferramentas software Matlab e Weka, así como tamén as técnicas de avaliación e as medidas de rendemento empregadas.

- **Capitulo 3: Selección de características e clasificación.** Neste capítulo preséntase unha visión global da selección de características, a súa definición e as principais vantaxes que aporta á aprendizaxe máquina, amosando os diferentes métodos de selección de características existentes, de igual xeito que se presentan os algoritmos de clasificación que se usan neste proxecto. Tamén se describe o proceso de discretización, necesario para que algúns dos métodos de selección poidan ser executados.

- **Capitulo 4: Experimentación e resultados.** Esta sección constitúe o eixe deste traballo, no que se amosan os resultados dos experimentos levados a cabo, analizando de forma exhaustiva os resultados obtidos que nos permitan chegar ás respostas que buscamos sobre as cuestións plantexadas.

- **Capitulo 5: Conclusións.** Nesta sección amosaranse as principais ideas acadadas a través da análise dos resultados dos distintos experimentos levados a cabo, e finalmente coméntanse algunhas das liñas de traballo futuro sobre as que se pode seguir traballando na materia.

Capítulo 2

Materiais e métodos

Neste capítulo da memoria presentaranse as diferentes ferramentas software que se empregaron neste proxecto xunto cunha breve descrición das mesmas, os distintos conxuntos de datos utilizados e as técnicas de avaliación dos resultados.

2.1 Ferramentas software

Para a realización da parte experimental correspondente a este proxecto empregáronse dúas ferramentas software:

- MATLAB R2019a
- Weka

2.1.1 MATLAB R2019a

Matlab é un programa de cálculo numérico, orientado a matrices e vectores, amplamente coñecido e empregado especialmente no ámbito científico e da enxeñaría. Foi creado en 1984 por TheMathWorks, e o seu nome procede da abreviatura Matrix Laboratory (Laboratorio de matrices).

MATLAB dispón na actualidade dun gran número de programas especializados que permiten ampliar as súas capacidades. Son as librerías especializadas ou toolboxes, que amplían de forma significativa o número de funcións incluídas no programa principal. Estas librerías cobren na actualidade practicamente todas as principais áreas no mundo da enxeñaría e da simulación, destacando os toolboxes de procesado de imaxes, estatística, análise financieira, matemáticas simbólicas, redes neuronais, lóxica difusa, etc. Trátase dun entorno de cálculo técnico, que se converteu en estándar da industria, con capacidades que todavía non foron superadas tanto en computación como en visualización numérica.

No caso deste proxecto fin de carreira, a utilización de MATLAB (a versión empregada é R2019b), é necesaria polo alto grado de compatibilidade coa ferramenta Weka necesaria para as distintas probas realizadas. Deste xeito, fixéronse programas en MATLAB desde os cales se preprocesan os datos, se preparan validacións cruzadas e se fan as chamadas oportunas a Weka.

2.1.2 Weka

Weka é unha colección de algoritmos de aprendizaxe máquina para tarefas de minería de datos. Trátase dun software de código aberto emitido baixo a Licencia Pública Xeral de GNU e contén ferramentas para a preparación de datos, clasificación, regresión, agrupación, minería de regras de asociación e visualización. Mediante esta aplicación temos a posibilidade de chamar aos métodos que desexemos empregar ou ben por interface gráfica ou por outra banda mediante comandos Java, invocados dende o código do programa. Da ferramenta anteriormente descrita empregáronse métodos para a selección de características como son Correlation-Based Feature Selection (CFS), InfoGain ou ReliefF.

2.2 Conxuntos de datos

As probas experimentais que se realizaron ao longo deste proxecto empregaron un amplo conxuntos de datos, que podemos separar en dous grupos según o número de características que presentan. Por un lado contamos con 49 bases de datos normais e 7 microarrays. A continuación presentase un breve resumo da natureza de ditos conxuntos de datos.

O primeiro bloque comprende 49 datasets extraídos dun repositorio amplamente utilizado por investigadores en aprendizaxe máquina, o repositorio Irvine da Universidade de California[1]. A maioría destes datasets presentan as condicións máis habituais, que son ter máis mostras ca características, aínda que hai algunha excepción, como por exemplo o dataset *arcene*, que conta con moitas máis características ca mostras e se tratará experimentalmente coma os datasets microarray que se describen a continuación. A Táboa 2.1 reflicte as características principais deste grupo de datasets, incluindo número de mostras, atributos e clases.

O segundo grupo de datasets está formado por sete datasets de tipo microarray [1, 2]. A clasificación de datos microarray plantexa un serio desafío para as técnicas computacionais debido á súa gran dimensión (miles de características/xenes) con tamaños mostrais pequenos. Ademais, existen complicacións adicionais que convirten a análise destes conxuntos nun dominio apasionante para os investigadores, como o desbalanceo das clases, ou presencia de valores atípicos (outliers). A Táboa 2.2 reflicte o número de mostras, atributos e clases dos datasets microarray escollidos para este estudo.

[1] https://archive.ics.uci.edu/ml/index.php

Conxunto de datos	# mostras	# atributos	# clases
arcene	200	10000	2
arrhythmia	452	279	13
basehock	1993	4862	2
bc-wisc-diag	569	30	2
bc-wisc-prog	198	33	2
breast	569	30	2
carcinom	174	9182	11
COIL20	1440	1024	20
congress	435	16	2
connbenchsonarminesrocks	208	60	2
connect4	67557	42	3
dermatology	358	34	6
gisette	7000	5000	2
glass	214	8	6
heart	270	13	2
hillvalley	606	100	2
ionosphere	351	34	2
isolet	7797	617	26
krvskp	3196	36	2
landsat	6435	36	6
libras	360	90	15
lowresspectR	531	100	9
molecbiolpromoter	106	57	2
molecbiolsplice	3190	60	3
musk2	6598	166	2
nci9	60	9712	9
optdigits	5620	64	10
orlraws10P	100	10304	10
ozone	2536	72	2
parkinsons	195	22	2
page_blocks	5473	10	5
PCMAC	1943	3289	2
pendigits	10992	16	10
pixraw10P	100	10000	10
RELATHE	1427	4322	2
satimage	6435	36	6
segmentation	2310	19	7
semeion	1593	256	10
sonar	208	60	2
soybeansmall	47	35	4
spect	267	22	2
splice	3175	60	3
USPS	9298	256	10
warpAR10P	130	2400	10
warpPIE10P	210	2420	10
waveform	5000	40	3
wine	178	13	3
yale	165	1024	15
zoo	101	17	7

Táboa 2.1: Características dos datasets

Conxunto de datos	# mostras	# atributos	# clases
9_Tumors	60	5726	9
CNS	60	7129	2
colon	62	2000	2
DLBCL	47	4026	2
Leukemia_1	72	5327	3
SRBCT	83	2308	4
TOX_111	171	5748	4

Táboa 2.2: Características dos microarrays

2.3 Técnicas de avaliación

Nesta sección analízase o deseño de experimentos de aprendizaxe automática, para avaliar e comparar o rendemento dos algoritmos de aprendizaxe na práctica e os test estatísticos, para analizar os resultados destes experimentos.

Ao avaliar e comparar algoritmos de aprendizaxe, é necesario dividir os conxuntos de datos en polo menos dous subconxuntos diferentes, chamados subconxunto de adestramento e subconxunto de proba, que estarán compostos por unha determinada porcentaxe de mostras do conxunto de datos. As mostras que forman parte do conxunto de probas non deben usarse no proceso de adestramento do algoritmo de aprendizaxe.

A taxa de erro do conxunto de adestramento, por definición, xeralmente é inferior á taxa de erro no conxunto de probas que contén mostras que non se usaron no adestramento. Polo tanto, as decisións non se poden tomar en función da taxa de erro de adestramento, como pode ser na comparación de dous algoritmos. Isto é debido a que sobre o conxunto de adestramento, o modelo máis complexo con máis parámetros case sempre produce menos erros que o modelo máis sinxelo.

Partindo da necesidade de dividir o conxunto de datos para obter o subconxunto de adestramento e o subconxunto de proba, hai varias estratexias de validación dispoñibles para un sistema de aprendizaxe dependendo da forma en que se divide o conxunto. Entre elas pódense destacar as seguintes:

- **Validación simple (hold-out validation)** [3, 4]. Este enfoque usa un conxunto de mostras para construír o modelo do clasificador e un conxunto diferente para estimar o erro, co fin de eliminar o efecto da sobreespecialización. Entre a variedade de porcentaxes empregadas, unha das máis frecuentes é tomar 2/3 das mostras para o proceso de aprendizaxe e o 1/3 restante para avaliar o seu rendemento. O feito de que só unha parte das mostras dispoñibles se empregue para levar a cabo a aprendizaxe é o principal inconveniente desta técnica, tendo en conta que se perde información útil no proceso de

indución do clasificador. Esta situación deteriora se o número de mostras para construír o modelo é moi pequeno, xa sexa pola porcentaxe escollida ou porque non se dispón de máis datos. Non obstante, hai conxuntos estándar (benchmarks) que xa están divididos en conxuntos de adestramento e conxuntos de probas e que foron usados deste xeito por diferentes autores, polo que é común empregar esta división.

- **Validación cruzada con k particiones (k-fold cross validation)** [3, 4]. Esta técnica proponse evitar a ocultación de parte das mostras do algoritmo de indución e a conseguinte perda de información. Con esta técnica o conxunto de datos divídese en k particións mutuamente excluíntes, que conteñen aproximadamente o mesmo número de mostras. Lévanse a cabo k validacións e en cada unha delas déixase un dos subconxuntos para a proba, e o sistema adestra co restante k-1. Así, a precisión estimada é a media das k precisións obtidas. Neste caso, os subconxuntos de proba son independentes, non sendo así os subconxuntos de adestramento. Algúns autores concluíron que os mellores resultados obtéñense cun valor de k = 10. Un caso particular deste método de avaliación é a validación cruzada sen saída, onde k é igual ao número de mostras do conxunto de datos. Neste caso, o clasificador adestra con todas as mostras menos unha para a súa proba. O principal inconveniente deste método é o elevado custo computacional da aprendizaxe do clasificador k veces, polo que non adoita utilizarse cando o número de mostras é alto ou o proceso de indución do clasificador é custoso.

2.4 Tests estatísticos

Cando se realizan múltiples execucións dun mesmo modelo obtéñense resultados diferentes, como por exemplo, despois de aplicar unha validación cruzada con k particións (k-fold cross validation). Nesta situación na que nos interesa avaliar a actuación de diferentes algoritmos sobre múltiples datasets, os test estatísticos permítennos comprobar a existencia de diferencias significativas respecto á media de cada uno de los modelos. Para explorar o significado estatístico dos resultados de clasificación no noso proxecto, usamos primeiro un test non paramétrico de Friedman [5] e un procedemento de comparación múltiple de Nemenyi [6]. O primeiro deles permite coñecer a existencia de diferenzas significativas entre varias mostras, pero non proporciona información sobre a mostra. Para coñecer esta información aplícase un segundo test, que neste caso é o procedemento de comparación múltiple mencionado anteriormente, e preséntanse os correspondentes diagramas de diferenza crítica, introducidos por Demšar [7], onde están conectados grupos de métodos que non son significativamente diferentes. Polo tanto, o proceso a seguir para realizar unha análise estatística dos modelos a comparar é:

- **Paso 1.** Aplícase o test non paramétrico de Friedman para comprobar se, para un nivel

de significación α, existen diferencias estatísticamente significativas entre as medias de cada un dos métodos que se probaron. Neste traballo utilizouse un valor de α igual a 0.05, é dicir, fíxose o test para un nivel de confianza do 95%.

- **Paso 2.** Se o test non paramétrico de Friedman é significativo, pódese afirmar que polo menos existe un modelo que ten un rendemento significativamente diferente ao resto. Neste caso é necesario aplicar un procedemento de comparación múltiple para obter os modelos que presentan o rendemento medio máximo.

- **Paso 3.** Finalmente, obtense o conxunto de modelos cuxos errores medios non son significativamente distintos entre eles, e represéntanse por medio de diagramas de diferencias críticas.

Capítulo 3

Selección de características e clasificación

Nos últimos anos desenvolvéronse novas aplicacións que manexan grandes cantidades de datos, tales como a minería de datos, o recoñecemento de patróns e a aprendizaxe máquina. Teoricamente, parece lóxico supoñer que contar con máis atributos ou características daríalle un maior poder de decisión a un algoritmo, mais isto non é así para os algoritmos de indución, xa que moitos sofren a *maldición da dimensión* [8, 9]. Este termo foi cuñado para nomear o fenómeno que se produce cando ao aumentarmos o número de características nunha tarefa, o tempo requirido polo algoritmo de indución para realizar o adestramento medra desmesuradamente, moitas veces de xeito exponencial [10]. Este problema, ademais de tempos de execución moi elevados, presenta outros inconvenientes como a aparición de atributos reduntantes e/ou irrelevantes [11]. Por todo isto, limitarmos o número de características que deben ser tratadas mantendo o rendemento converteuse nun problema crucial.

Fundamentalmente, pódense distinguir dúas aproximacións para reducir o número de características dun problema dado: a selección e a extracción de características. A selección trata de escoller do conxunto orixinal un subconxunto de características que conteña suficiente información para abordar un problema determinado [12]. Á súa vez, a extracción crea un subconxunto de características novas mediante a combinación das características existentes [13].

A selección de características pode aplicarse a problemas tanto de regresión como de clasificación. Non obstante, a maioría dos estudos neste campo céntranse en problemas de clasificación [14, 15, 16], aínda que a súa aplicación a problemas de regresión está estendéndose [17, 18, 19, 20].

Neste traballo fin de grao utilizarase a selección como técnica para reducir a dimensión do espazo de características. Así mesmo centrarémonos no problema da clasificación co propósito de poder realizarmos un estudo comparativo cos resultados derivados doutras aproxi-

macións. Además, nesta sección tamén se describirá brevemente o proceso de discretización, necesario como paso previo á execución dalgún dos métodos de selección de características que se van estudar.

3.1 Selección de características

A selección de características relevantes débese á preferencia polos modelos máis sinxelos fronte aos máis complexos e ten as súas orixes no coñecido *principio da navalla de Occam* [11, 21, 22]. Existen múltiples definicións para explicar que é, algunhas delas intuitivas, como é o caso da proporcionada por Hall [23]:

Definición 1 *A **selección de características** é o proceso de identificar e eliminar tanta información irrelevante e redundante como for posible.*

Polo tanto entendemos que o principal obxectivo da selección de características é poder obtermos un subconxunto de características que describa axeitadamente o problema dado, de maneira que este subconxunto sexa o empregado durante o proceso de adestramento.

Realizar a selección das características relevantes e poder eliminar aquelas das que non se pode extraer coñecemento supón unha serie de beneficios importantes, como poden ser os seguintes [11, 24, 25]:

- As características irrelevantes e redundantes poden confundir os algoritmos de aprendizaxe, polo cal de se eliminaren as ditas características o clasificador obtido será máis exacto.
- As características que describen as mostras determinan o espazo de busca que debe explorar o algoritmo de aprendizaxe. Así, reducindo a dimensión dos datos redúcese o tamaño do espazo de busca, o que lles permite aos algoritmos operaren máis rápido e eficientemente.
- Unha vez que se demostra que algunhas características son prescindibles, é innecesario gardalas e, por tanto, redúcense os requisitos de almacenamento de información.
- A redución de características poderíase ter en conta en futuras recollidas de datos, pois sería posible que implicase unha redución nos custos económicos. Un exemplo témolo nos diagnósticos clínicos, en que as mostras están formadas tanto por síntomas observables como polos resultados de probas médicas. Estas probas teñen asociados certos riscos e custos, como poden ser os relativos a unha exploración cirúrxica invasiva. O feito de sinalarmos que unha proba deste tipo non é necesaria para determinar un diagnóstico supón unha diminución nos custos considerable.

- Fai máis doada a comprensión dos datos e a súa visualización. Ao eliminarmos as características redundantes e irrelevantes o clasificador obtido será máis sixelo, o que vai facilitar a comprensión dos resultados.

- Reduce o risco de sobreaxuste. Este problema prodúcese cando o algoritmo de indución se axusta demasiado aos datos de adestramento, é dicir, acada unha precisión moi elevada durante o proceso de adestramento, mais a predición obtida con datos novos é moi baixa [10]. Para evitar o sobreaxuste cando se dispón dun gran número de características, normalmente requírese un elevado número de mostras. Porén, obtermos un gran número de mostras non é factible en moitos problemas reais e, xa que logo, o risco de sobreaxuste redúcese canto menor é o número de características empregadas.

3.1.1 Relevancia das características

Da sección anterior pódese deducir que un dos obxectivos principais da selección de características é escoller as características relevantes, motivo polo cal se fai necesario aclarar que entendemos por relevancia.

Na bibliografía referida á aprendizaxe máquina non existe consenso sobre unha definición única do concepto de relevancia. A razón para esta diversidade é que, como indica Blum [17], todo depende polo xeral da pregunta: «Relevante para que?». Isto indica que segundo for o obxectivo pode ser máis apropiada unha ou outra definición.

Considérese un escenario no cal existen n características ou atributos empregados para describiren as mostras ou os exemplos dun determinado problema. Cada característica i pertence a un dominio F_i (dominio da característica X_i) e, por exemplo, unha característica pode ser lóxica, discreta con múltiples valores ou continua. O algoritmo de aprendizaxe recibirá como entrada un conxunto $\mathcal{D}$ de datos de adestramento onde cada exemplo ou mostra será un par formado por unha combinación de valores para os distintos atributos e unha saída ou etiqueta de clase.

De forma intuitiva poderíase determinar que unha característica é relevante cando achega información á resolución dun problema dado. Non obstante, esta é unha definición demasiado superficial, polo que se recorreremos á bibliografía existente para poder ofrecermos unha definición máis formal [18].

Almuallim e Dietterich [26] definen a relevancia baixo a asunción de que todas as características e as etiquetas de clase son binarias e de que non existe ruído.

Definición 2 *Unha característica X_i é **relevante** respecto a un concepto $\mathcal{C}$ se X_i aparece en calquera fórmula binaria que represente $\mathcal{C}$ e en caso contrario é irrelevante.*

Gennari *et al.* [27] permiten datos con ruído e características multivaluadas:

Definición 3 *X_i é **relevante** se e só se existen algúns valores x_i, e con $p(X_i = x_i) > 0$, tales que*

$$p(Y = y|X_i = x_i) \neq p(Y = y)$$

onde *Y* é a etiqueta da clase.

Conforme esta definición, X_i é relevante se unha vez coñecido o seu valor pode cambiar a estimación da etiqueta de clase *Y*. Nótese que esta definición falla ao capturarmos a relevancia de características no concepto de paridade onde todas as mostras sen etiquetar son equiprobables, e pode ser modificada como segue:

Sexa $S_i = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$ o conxunto de todas as características agás X_i. Defínese s_i como unha asignación simultánea de valores a todas as características de S_i.

Definición 4 *X_i é **relevante** se e só se existen algúns valores x_i, y, s_i con $p(X_i = x_i) > 0$, tales que*

$$p(Y = y, S_i = s_i|X_i = x_i) \neq p(Y = y, S_i = s_i)$$

Con esta definición, X_i é relevante se a probabilidade da etiqueta (dadas todas as características) pode cambiar cando eliminamos o coñecemento sobre o seu valor.

Algúns autores [17] propoñen outras definicións nun marco máis xeral:

Definición 5 (Relevancia respecto ao obxectivo) *Unha característica X_i é relevante respecto a unha función obxectivo c se existe un par de exemplos A e B no espazo de mostras tal que A e B difiren só no seu valor de X_i e $c(A) \neq c(B)$.*

Noutras palabras, unha característica X_i é relevante se existe algún exemplo no espazo de mostras para o cal o cambio do valor de X_i lle afecta ao valor da etiqueta. Agora ben, esta definición presenta o inconveniente de que o algoritmo de aprendizaxe, tendo só acceso ao conxunto de mostras $\mathcal{D}$, non ten por que determinar necesariamente se a característica X_i é relevante ou non. Aínda peor, se a codificación de atributos é redundante (atributos que conteñen a mesma información), pode incluso non ser posible que dúas mostras difiran soamente nun atributo.

John *et al.* [18] consideran que son necesarios dous graos de relevancia, feble e forte, para o cal a relevancia debería ser definida en termos dun clasificador de Bayes (o clasificador óptimo para un problema dado). Unha característica X_i presenta **relevancia forte** se eliminarmos X_i produce unha deterioración no rendemento dun clasificador de Bayes óptimo. Unha característica X_i presenta **relevancia feble** se non presenta relevancia forte e existe un conxunto de características S tal que o rendemento dun clasificador de Bayes sobre S é peor que o rendemento sobre $S \cup \{X_i\}$. Unha característica que non presenta nin relevancia forte nin feble é, daquela, **irrelevante**.

Definición 6 (Relevancia forte) *Unha característica X_i é **fortemente relevante** se e só se existe algún x_i, y, s_i para o cal $p(X_i = x_i, S_i = s_i) > 0$ tal que*

$$p(Y = y|X_i = x_i, S_i = s_i) \neq p(Y = y|S_i = s_i)$$

Definición 7 (Relevancia feble) *Unha característica X_i é **debilmente relevante** se e só se non é fortemente relevante, e hai un subconxunto de características $S'_i \subseteq S_i$, para o cal existen algúns valores x_i, y, s'_i con $p(X_i = x_i, S'_i = s'_i) > 0$, tales que*

$$p(Y = y|X_i = x_i, S'_i = s'_i) \neq p(Y = y|S'_i = s'_i)$$

É dicir, unha característica é debilmente relevante cando, eliminando un subconxunto de características do espazo de entrada, esta se converte en fortemente relevante.

Estas definicións de relevancia son útiles desde o punto de vista dun algoritmo de aprendizaxe á hora de decidir que características manter e cales ignorar. Xeralmente é importante mantermos as características fortemente relevantes, cando menos no sentido de que eliminando unha característica con relevancia forte engádeselle ambigüidade ao conxunto de mostras. O feito de manter ou non as características debilmente relevantes depende de que características fosen xa ignoradas.

En moitos casos, máis do que estar preocupados con que características son relevantes, simplemente quere utilizarse a relevancia como unha *medida de complexidade*. É dicir, quérese utilizar a relevancia para indicar como de "complicada" é unha función, e máis que pedir que o algoritmo de aprendizaxe seleccione un subconxunto de características, apenas se pretende que o seu rendemento sexa satisfactorio cando a cantidade de características é pequena. Para este propósito proponse unha nova definición de relevancia como unha medida de complexidade con respecto a un conxunto de mostras $\mathcal{D}$ e a un concepto $\mathcal{C}$:

Definición 8 (Relevancia como medida de complexidade) *Dado un conxunto de mostras $\mathcal{D}$ e un conxunto de concepto $\mathcal{C}$, sexa r($\mathcal{D}, \mathcal{C}$) o número de características relevantes para o concepto en $\mathcal{C}$ usando a Definición 5, de tal xeito que se acade a maior precisión utilizando o menor número de características relevantes.*

Noutras palabras, preténdese atopar o menor número de características necesario para acadarmos un funcionamento óptimo no conxunto de aprendizaxe $\mathcal{D}$ a través do concepto $\mathcal{C}$. A razón pola cal se especifica o concepto $\mathcal{C}$ é que pode que exista unha característica, como pode ser o número da seguridade social, que ten moita relevancia desde o punto de vista da información contida, mais pode ser totalmente inútil á hora de utilizala para a clasificación de certos conceptos.

As definicións vistas ata o momento son independentes do algoritmo de aprendizaxe específico que está sendo empregado. Segundo estas definicións non existe garantía de que unha característica, por ser relevante, lle sexa necesariamente útil ao algoritmo de aprendizaxe, ou de que unha característica irrelevante non lle sexa útil. Por isto cómpre falarmos dun novo concepto: *optimalidade* ou *utilidade* . En Caruana e Freitag [28] fan explícito este feito co que eles chaman "utilidade" :

Definición 9 (Utilidade incremental) *Dado un conxunto de mostras $\mathcal{D}$, un algoritmo de aprendizaxe $\mathcal{I}$ e un conxunto de características $\mathcal{S}$, unha característica X_i é incrementalmente útil con respecto a $\mathcal{S}$ se a precisión que acada $\mathcal{I}$ empregando o conxunto de características $\{X_i\} \cup \mathcal{S}$ é maior que a precisión obtida empregando soamente o conxunto $\mathcal{A}$.*

Un concepto relacionado coa relevancia é a *redundancia* entre características, que se expresa normalmente en termos de correlación entre elas. Segundo Guyon e Elisseeff [24] considérase que dúas características son redundantes se os seus valores están completamente "correlacionados", no sentido de que non hai ganancia de información cando ambas son engadidas. Diversos estudos demostraron que o subconxunto de características óptimo é o formado polas características fortemente relevantes e as características debilmente relevantes non-redundantes [14].

3.1.2 Interacción entre características

Enténdese que dúas características interaccionan se unha característica que por si soa non é útil pode proporcionar unha mellora significativa no rendemento dun clasificador cando se considera con outras [16, 24]. Jakulin [29, 30] cuñou o termo *interacción positiva* para facer referencia a estas relacións. De facermos caso omiso das interaccións entre características non se está considerando toda a información que nos proporciona o conxunto de datos.

Considérese o problema do OR exclusivo $Y = XOR(X_1, X_2)$, onde Y é unha clase binaria e X_1 e X_2 son atributos binarios. Se observamos o atributo X_1 individualmente, este non proporciona unha evidencia sobre os valores de Y. A razón é que a relación entre X_1 e Y depende de X_2. Para $X_2 = 0, Y = X_1$; para $X_2 = 1, Y \neq X_1$. O caso é análogo se consideramos X_2 de forma individual. No entanto, X_1 e X_2 xuntos determinan Y perfectamente. Dise que existe unha interacción positiva entre X_1 e X_2 con respecto a Y. No caso de que se produza unha interacción positiva a evidencia de considerarmos X_1 e X_2 xuntos é maior que a suma da evidencia de X_1 e a evidencia de X_2 consideradas de forma individual.

Aquí cabe destacar que, aínda que no exemplo anterior se trata a interacción entre dúas características, esta non está limitada unicamente a relacións dúas a dúas, senón que poden entrar en xogo tres ou máis características.

É importante que un método de selección de características contemple a interacción entre elas, dado que, se asumirmos que son independentes, pódense obter solucións pouco acertadas dos problemas propostos, pois características altamente relevantes vense descartadas.

Non obstante, non existen demasiados métodos de selección que contemplen a interacción de características. Dentro dos métodos de filtrado cabe destacar o INTERACT [16], que recibe o seu nome precisamente pola súa capacidade para detectar a interacción. Con respecto aos métodos envolventes, dependerá do algoritmo de indución empregado. Así, o algoritmo naive Bayes [31] non considera a interacción entre características, pero o algoritmo C4.5 [32] si a ten en conta.

3.1.3 Métodos de selección

Como indicamos en puntos anteriores, hai varios algoritmos de selección de características que abordan o problema dende diferentes puntos de vista. A pesar disto, todos comparten a mesma definición formal do problema e pódense clasificar en familias numerosas segundo (a) a interacción entre o proceso de indución e selección e (b) o proceso de busca implementado. Por iso é útil comezar cun percorrido por estes aspectos, xa que nos ofrece un marco común no que enmarcar os algoritmos implementados neste traballo.

3.1.3.1 Compoñentes dun método de selección de características

Xa que un algoritmo de selección de características busca un conxunto óptimo de funcións para un problema dado, este presenta os mesmos elementos que un proceso de busca, é dicir [33]:

- **Estado inicial.** O seu valor dependerá dos operadores que xeren os estados sucesores e da dirección da busca. No caso de que a busca sexa cara atrás (backward), comeza co conxunto completo de características e estas eliminaranse paso a paso. Se a busca é cara adiante (forward), pártese o conxunto baleiro e engádense características progresivamente. Outra alternativa sería a busca aleatoria, para a que se parte dun subconxunto aleatorio de características.

- **Espazo de estados.** Formado por todos os posibles subconxuntos de características. Cada estado do espazo de estados indica unha solución concreta, é dicir, cales son as características que deberían ser seleccionadas. Para un conxunto de datos con n características, o número de posibles subconxuntos sería $2n$. Obviamente, na maioría dos casos a realización dunha busca exhaustiva no espazo de estados é impracticable, polo que se fai necesario empregar una estratexia de busca adecuada que realice una exploración dirixida.

- **Estratexia de busca.** Como se viu no epígrafe anterior, non é viable abordar unha búsqueda exhaustiva e, por tanto, soen empregarse diversas estratexias de busca coa finalidade de reducir o tempo de execución. Algunas destas estratexias son as coñecidas como ascensión a colinas (hill climbing) [34] ou primeiro o mellor (best first) [34, 35].

- **Función de avaliación.** Determina a calidade de cada conxunto de características. Esta función pode ser de varios tipos e incluir medidas de información, distancia, dependencia, consistencia ou exactitude [36].

- **Criterio de parada.** Establece a condición que se debe cumprir para que o algoritmo deixe de iterar. Por exemplo, o proceso remata se a función de evaluación non mellora despois de agregar ou eliminar unha nova característica.

3.1.4 Clasificación dos métodos de selección de características

Existen numerosos estudos sobre diversos aspectos da selección de características que agrupan os distintos algoritmos existentes na literatura. Una destas clasificacions é a establecida por Blum e Langley [17] na que, ademais de ter en conta a función de avaliación, tamén empregan como criterio a dependencia que existe entre o proceso de selección e o de indución e agrupan os métodos en tres categorías (ver Figura 3.1, que inclúe as vantaxas e inconvenientes de cada método.).

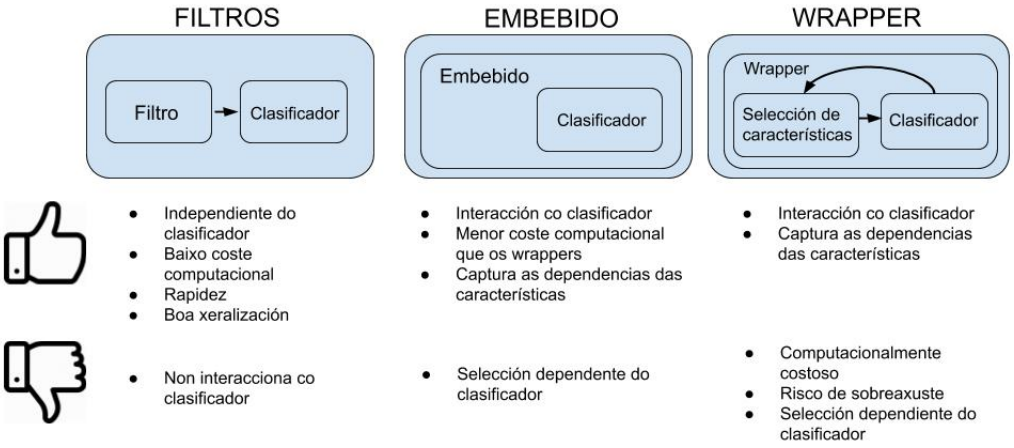

Figura 3.1: Tipos de métodos de selección de características

Na primeira categoría inclúense os *métodos encaixados* (*embedded*), en que o indutor conta co seu propio algoritmo de selección de características, como ocorre nos algoritmos que xeran árbores de decisión, os cales usan só aqueles atributos necesarios para obter unha descrición

consistente co conxunto de aprendizaxe. É o máis simple dos métodos, e o menos empregado. Alén das árbores de decisión para a clasificación e regresión (CART, *Classification and Regression Tree*), outro exemplo de métodos encaixados son as redes de neuronas artificiais (RNA).

Na segunda categoría, os *filtros*, inclúense os algoritmos en que a selección de características se realiza como un paso previo á fase de indución e, por tanto, de maneira independente a esta, baseándose en propiedades dos propios datos como poden ser a correlación coa clase, polo que pode entenderse como un filtrado dos atributos (ou características) irrelevantes e redundantes.

Nos *métodos envolventes* (*wrappers*), a selección de atributos e os algoritmos de aprendizaxe non son elementos independentes, xa que a selección dos atributos fai uso do proceso de indución para avaliar a calidade de cada conxunto de atributos seleccionados en cada momento. A principal diferenza entre estes dous métodos reside en que os filtros se basean nas características xerais dos conxuntos de adestramento para realizaren a selección, mentres que os *wrappers* precisan dun algoritmo de aprendizaxe predeterminado para identificaren as características que deben ser seleccionadas.

Na actualidade desenvolvéronse novos métodos que poderían clasificarse como métodos híbridos [15] que, principalmente, combinan as vantaxes dos filtros e dos métodos envolventes intentando eliminar as suas limitacións co obxectivo de alcanzar un bo rendemento nun tempo razoable.

3.1.5 Algoritmos de filtrado

Nesta sección realizarase unha descrición dos aspectos más importantes dos filtros por seren esta a aproximación máis empregada na actualidade e sobre a que se baseará o desenvolvemento do presente traballo.

Neste tipo de métodos, a selección de características é tratada coma un paso previo á extracción de coñecemento, i.e., trátase de prover unicamente das características que máis información conteñan ao proceso de inducción, de aí que John *et al.* [18] os denominasen filtros. Segundo Blum [17], este paso de preprocesado emprega as propiedades inherentes dos datos de entrada para seleccionar algunhas características e excluir outras. Debido a isto, os métodos de filtrado son independentes do algoritmo de indución, que se limitará a tomar a saída proporcionada polo filtro. O esquema máis simple de filtro consiste en avaliar cada característica individualmente baseándose, por exemplo, na correlación coa etiqueta de clase. Unha vez avaliadas todas as características por separado, selecciónanse as k características cos valores de correlación máis altos. Non obstante, existen variantes máis complexas nas que se avalían múltiples características simultáneamente para analizar a sua interacción.

Dentro dos filtros, podemos distinguir dous enfoques en función da maneira na que elixen

as características:

- **Avaliación individual** (*rankers*). Avalían cada unha das características dando lugar a unha clasificación ordenada de todas as características. Unha vez ordenadas, elíxense as k mellores, sendo k un parámetro determinado polo usuario.
- **Avaliación de subconxunto**. Avalían subconxuntos de características e dan como resultado o subconxunto que consideran que aporta a maior información. As características pertencentes ao subconxunto resultado non están ordeadas, polo que non é posible limitar o seu número por parte do usuario.

Entre as ventaxas que presentan estos algoritmos, cabe destacar as seguintes [37, 38, 39]:

- **Velocidade de execución.** O feito de non ter que realizar sucesivos adestramentos dun algoritmo de indución sobre o conxunto de datos supón un menor custo computacional e, por tanto, un menor tempo de execución. Non obstante, algunhas medidas de correlación son moi custosas de calcular.
- **Aplicabilidade a conxuntos de gran tamaño.** Gracias á sua maior rapidez, poden ser aplicados a grandes conxuntos de datos.
- **Evitación do sobreaxuste.** Naqueles conxuntos onde o número de mostras é reducido e o número de características elevado, os algoritmos de indución tenderán a sobreaxustar os datos. A eliminación de características por parte do filtro evita este sobreaxuste.

Por outra parte, este tipo de métodos tamén presentan unha serie de inconvenientes, entre os que se poden destacar a selección de subconxuntos de características demasiado grandes, en especial nos filtros de subconxunto, xa que non é posible limitar o número de características seleccionadas. Por outra parte, amosan un baixo rendemento teórico debido a que non teñen en conta a precisión de clasificación tras realizar a selección. Isto tamén permite, nalgúns casos, unha maior xeralización, xa que a selección de características non depende de ningún clasificador en concreto. Non obstante, os resultados empíricos mostran que certos criterios obteñen moi bos resultados.

No Algoritmo 1 amósase o pseudocódigo dun algoritmo de filtrado xenérico. Para un conxunto de datos de partida D, o algoritmo comeza a búsqueda partindo dun subconxunto de características S_0, que pode ser o conxunto completo de características, o conxunto vacío ou calquera subconxunto aleatorio do conxunto completo de características. A continuación, faise a búsqueda a través do espazo de estados baixo unha estratexia de búsqueda particular ata alcanzar o criterio de parada δ. Sobre cada subconxunto S aplícase unha función de avaliación M independente do algoritmo de indución e compárase o resultado co mellor subconxunto obtido ata o momento. En caso de que o conxunto S sexa mellor que S_*, almacénase S como novo subconxunto óptimo.

Entrada: $\mathcal{D}(X_1, X_2, \ldots, X_n)$ ▷ Conxunto de adestramento con *n* características
Entrada: S_0 ▷ Subconxunto desde o que se comeza a busca (estado inicial)
Entrada: δ ▷ Criterio de parada
Saída: S_{opt} ▷ Subconxunto óptimo de características

1. **Dáse un valor inicial** $S_{opt} = S_0$
2. $\gamma_{mellor} = avaliar(S_0, \mathcal{D}, M)$ ▷ Avalíase S_0 mediante unha función de avaliación *M*
3. **Repítese**
   - 3.1 $S = xerar(\mathcal{D})$ ▷ Xérase un subconxunto para a súa avaliación
   - 3.2 $\gamma = avaliar(S, \mathcal{D}, M)$ ▷ Avalíase o actual subconxunto mediante *M*
   - 3.3 **Se** γ é mellor que γ_{mellor} **entón**
     - 3.3.1 $\gamma_{mellor} = \gamma$
     - 3.3.2 $S_{opt} = S$
   - 3.4 **Fin se**
4. **Ata que** (δ é acadado)
5. **Devolve** S_{opt}

Algorithm 1: Algoritmo xenérico de filtrado

3.1.6 Algoritmos de filtrado empregados

Neste traballo vanse utilizar algoritmos de filtrado debido a que se traballa con conxuntos de datos de gran tamaño e a aplicación de *wrappers* non se mostra eficiente ante este tipo de problemas, xa que levaría demasiado tempo e consumiría demasiada memoria. Con respecto aos métodos híbridos, non serán empregados porque cos filtros xa se consegue unha gran redución do número de características e, xa que logo, non sería lóxico aplicar a continuación un *wrapper*.

Os filtros están formados polos seguintes compoñentes:

- **Algoritmo avaliador:** É a función que determina a calidade do conxunto de atributos para discriminar a clase.
- **Algoritmo de busca:** Realizará a exploración polo espazo de estados de todos os posibles subconxuntos de características.

Deste xeito, para obtermos distintos filtros abonda con combinarmos distintos algoritmos de avaliación e busca. Así pois, para levar a cabo o estudo comparativo escolléronse dous algoritmos de avaliación, *CFS* [23], e *INTERACT* [16], combinados cunha estratexia de busca 'primeiro o mellor' (ou *best-first*) [40, 41], e catro algoritmos de avaliación, *Information Gain*

[32, 42], *ReliefF* [43, 44] *MIM* [45] e *mRMR* [46], combinados cunha estratexia de busca de tipo 'ranker', que devolve unha lista de todos os atributos segundo a súa relevancia.

3.1.6.1 Correlation-Based Feature Selection (CFS)

O filtro CFS (*Correlation-based Feature Selection*), desenvolvido por M. A. Hall a finais dos 90 [23, 47], foi o primeiro método de selección baseado na análise de subconxuntos de características. A súa principal meta é obter o subconxunto de características máis correlacionado coa clase e menos correlacionado entre si.

A heurística empregada polo algoritmo de filtrado CFS ten en consideración a utilidade das características individuais para predicir a clase, ademais de estudar o nivel de intercorrelación entre elas. A medida empregada é a seguinte:

$$q_{zc} = \frac{n\overline{q_{zi}}}{\sqrt{n + n(n-1)\overline{q_{ii}}}} \tag{3.1}$$

Onde q_{zc} é a correlación entre as características individuais e a saída, n é o número de características, $\overline{q_{zi}}$ é a media das correlacións entre as características individuais e a variable de saída, e $\overline{q_{ii}}$ é a intercorrelación media entre compoñentes. Esta medida asígnalles valores altos aos subconxuntos fortemente correlacionados coa saída e debilmente correlacionados entre si.

As características irrelevantes deberían ser ignoradas polo mero feito de que terán baixa correlación coa saída. As características redundantes, á súa vez, tamén deberían ser descartadas porque estarán altamente correlacionadas con unha ou máis das restantes características. A aceptación ou non dunha característica dependerá, por tanto, de como as devanditas características predigan a saída en áreas do problema aínda non cubertas por outras características.

O numerador da ecuación (3.1) pode observarse como un indicador da capacidade de predición dun grupo de características, sendo o denominador un índice da redundancia entre elas dentro do conxunto. Como cada característica é tratada de forma individual, o algoritmo CFS non pode identificar interaccións realmente fortes como as que poderían aparecer en problemas de paridade. Pese a todo, este algoritmo ofrece bos resultados en problemas en que a interacción non for demasiado elevada.

A aplicación deste algoritmo require a realización de dous pasos. No primeiro de eles discretízanse as características numéricas e posteriormente emprégase unha medida denominada incerteza simétrica *SU, Symmetrical Uncertainty*, desenvolvida por Press *et al.* en 1988 [48], e que se define como o cociente entre o aumento da información e a entropía de dúas características A e B:

$$SU(A,B) = 2\left\{\frac{IG(A/B)}{H(A) + H(B)}\right\}. \tag{3.2}$$

Para comprendermos a ecuación anterior é necesario explicarmos as medidas que nela se tratan; así, contamos con:

- Entropía (H): É a medida da incerteza dunha característica escollida ao azar. Definida por diversos autores ao longo da historia, foi Shannon o que lle achegou a esta propiedade un sentido relacionado coa información en 1948 [45].

- Ganancia de información (*Information Gain (IG)*): Esta medida desenvolvida por John Ross Quinlan en 1993 [32] defínese mediante a seguinte ecuación:

$$IG(A|B) = H(A) - H(A|B) \tag{3.3}$$

  Esta ecuación vén a tratar cos valores $H(A)$, que representa a entropía da característica A, e $H(A/B)$ que contempla a entropía desta mesma característica unha vez que se observan os valores que toma outra característica B. O valor de IG indica a relevancia dunha característica con respecto a outras; así, no caso de que $IG(A/B) > IG(C/B)$ demostraríase que a característica B está máis correlacionada coa característica A do que coa característica C.

Se volvemos ao algoritmo CFS, emprégase a medida *SU*, ecuación (3.2), para estimarmos o grao de asociación entre as distintas características e obtermos así distintos subconxuntos. No segundo paso do algoritmo, emprégase a medida q_{zc}, ecuación (3.1), para seleccionarmos o subconxunto óptimo.

O filtro CFS non require especificamente que a medida empregada polo método sexa a *SU*, senón que tamén se poderían utilizar outras como a MDL (*Minimum Description Length*) ou o filtro Relief [49, 50].

3.1.6.2 INTERACT (INT)

O algoritmo INTERACT [16], desenvolvido por Z. Zhao e H. Liu, é un método baseado en medidas de inconsistencia e incerteza simétrica (SU, ecuación (3.2)) que introduce como notable mellora a detección da interacción entre características.

O motivo do desenvolvemento deste algoritmo baséase na idea de que, nun principio, unha característica pouco correlacionada coa saída podería parecer irrelevante, mais de se combinar con outras características podería chegar a estar moi correlacionada coa saída, sendo así unha característica importante para o estudo do conxunto. Eliminarmos inicialmente a devandita característica podería provocar unha mala selección final, pero tratar a interacción entre todas as características é prácticamente inviable, razón pola cal os algoritmos existentes ata o momento rara vez trataban este problema e adoitaban asumir a independencia entre todas as características.

Un exemplo claro de interacción entre variables amósase no problema denominado Monk1 [51]. Neste problema existen seis características booleanas e unha saída que se obtén como $(A_1 = A_2)|(A_5 = 1)$. Se se consideran individualmente, a correlación entre A_1 e a clase C (igualmente para A_2 e C) é cero, polo que A_1 e A_2 serían irrelevantes. Obsérvase, por tanto, que A_1 e A_2 só serán relevantes no caso de tratalas conxuntamente.

Os autores deste algoritmo consideran que a interacción entre características pode manexarse cun coidado deseño da medida de avaliación (*Consistency Contribution*) e a estratexia de busca (cara a atrás), deseñando tamén estruturas de datos acordes.

A principal medida empregada por este algoritmo, xunto co *SU* (ecuación (3.2)), é a *Consistency Contribution (c-contribution)*, que se refire a unha característica concreta e indica como lle afecta a eliminación desta característica á consistencia final do problema. O funcionamento do método divídese en dúas partes principais: na primeira delas realízase unha ordenación descendente das características con base no *SU* de cada unha delas, para posteriormente calcular a *c-contribution* de cada unha, comezando pola derradeira. Se o valor da *c-contribution* é menor que o valor establecido por un limiar, elimínase a característica, e se non, mantense.

3.1.6.3 Information Gain (IG)

Este filtro InfoGain (Information Gain) [42] é un dos métodos máis comúns de avaliación de atributos. É un filtro univariado que avalía as características segundo a súa ganancia de información (ver ecuación 3.3) e considera só unha característica á vez. Require que as variables numéricas sexan discretizadas. Ofrece unha clasificación ordenada de todas as características, e logo é necesario un limiar para manter un certo número delas na orde en que as ordenou o filtro. Este filtro usa a ganancia de información de cada característica con respecto á clase. A ganancia de información é unha medida usada habitualmente para a construción de árbores de decisión como método de clasificación, que mide a entropía ou trastorno no sistema segundo a Teoría da información [52].

3.1.6.4 ReliefF (RelF)

ReliefF [43] é unha extensión do algoritmo orixinal Relief. O Relief orixinal proposto por Kira e Rendell en [44], pretende estimar a *calidade* das características en función do ben que axudan a distinguir a clase entre instancias próximas no espazo das características. Só se aplica a problemas de clasificación binaria pero as características poden ser tanto nominais coma numéricas. Ten outra limitación importante, xa que non admite datos incompletos. Relief toma de forma aleatoria unha instancia dos datos para logo localizar o veciño máis próximo da mesma clase e o da clase contraria. Os valores característicos dos veciños máis próximos compáranse coa instancia tomada e úsanse para actualizar as puntuacións de relevancia para

cada característica. O razoamento é que unha característica útil debe ter valores diferentes para instancias de clases diferentes e valores similares para instancias dunha mesma clase.

Sexan $I1, I2$, calquera das duas instancias próximas, a idea intuitiva subxacente a Relief é a seguinte: unha característica C que toma valores diferentes para as instancias dadas $((C, I1) = (C, I2))$ é boa se as instancias son de diferente clase e malas se as instancias son da mesma clase. Nos dous casos, a diferenza no valor da característica axuda a separar as instancias no espazo de características, o que é desexable no primeiro caso pero non no segundo. A continuación, o vector de peso sería actualizado, o que indica a calidade de cada característica. Este proceso repetirase varias veces.

En [43], Kononenko revisa o algoritmo de Relief e propón varias extensións co fin de facelo máis robusto e superar algunhas das limitacións que ten. ReliefF engade a capacidade de tratar problemas con varias clases e tamén é capaz de tratar con datos e ruído incompletos. Este método pódese aplicar en todas as situacións, ten un sesgo baixo, inclúe a interacción entre as características e pode capturar dependencias locais das que outros métodos non son capaces.

A principal diferenza entre ReliefF e Relief é que agora, en vez de considerar só o veciño máis próximo á instancia analizado en cada iteración, son considerados os b veciños máis próximos. Outra diferenza é que considera varias clases no canto de só dúas. Tómanse os b veciños máis próximos de cada unha das c posibles clases (hai, polo tanto, $b(c-1)$ instancias que pertencen ás $c-1$ clases diferentes).

ReliefF proporciona como saída unha lista ordenada de características. Para quedarse cun certo número de características, neste traballo tómase como limiar o número de características que foron seleccionadas co filtro CFS, de igual forma que faremos con InfoGain, e seguen a orde establecida.

3.1.6.5 Maximización da Información Mutua (MIM)

Podemos definir a información mutua [45] entre X e Y ou, dito doutro xeito, a información que comparten X e Y, da seguinte forma:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}. \end{aligned} \tag{3.4}$$

Enténdese, polo tanto, que a información mutua entre X e Y é a diferenza entre a incerteza antes de coñecer Y, $H(X)$, e la incerteza despois de coñecer Y, $H(X|Y)$.

A correlación entre unha característica e a clase pode medirse a partir da información

mutua dunha característica x_i coa clase Y. A expresión deste criterio é

$$MIM(x_i) = I(x_i; Y)$$

e coñécese como Maximización da Información Mutua (en inglés: Mutual Information Minimization). Cabe destacar que este criterio, amplamente usado na literatura, considera todas as características de maneira independiente,xa que unicamente é necesario calcular o valor MIM para cada característica e escoller as k características con maior valor.

3.1.6.6 mínima Redundancia Máxima Relevancia (mRMR)

O método mRMR (minumum Redundancy Maximum Relevance) selecciona as características que posúen maior relevancia coa clase de obxectivo e tamén teñen unha redundancia mínima, é dicir, selecciona as características que son máis diferentes de cada unha das outras [46]. Ambos criterios (máxima relevancia e mínima redundancia) baséanse en información mutua.

En termos de información mutua, o obxectivo da selección de características é atopar un conxunto de características S con m atributos $\{x_i\}$, que xuntos proporcionan a maior dependencia sobre a clase obxectivo c. Este esquema chámase Dependencia Máxima (Max-Dependency) e identifícase coa seguinte fórmula:

$$\max D(S, c), \qquad D = I(\{x_i, i = 1, ..., m\}; c) \tag{3.5}$$

Obter o valor final de Máxima Dependencia adoita ser difícil, especialmente en problemas con conxuntos de datos formados por variables continuas. Incluso para variables discretas, non se poden evitar completamente os problemas do cálculo da Máxima Dependencia. Outro problema no cálculo deste valor é a baixa velocidade de computación.

Debido ás dificultades que presenta para calcular o valor da *Máxima Dependencia* dos conxuntos de datos, proponse unha alternativa, que consiste en seleccionar as características segundo a *Máxima Relevancia (Max-Relevance)*. A máxima relevancia consiste en seleccionar as características que satisfagan a ecuación (3.6), que aproxima o valor $D(S, c)$ presentado na ecuación (3.5) co valor medio de todos os valores de información mutua entre as características individuais x_i e a clase c:

$$\max D(S, c), \qquad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{3.6}$$

É probable que as características seleccionadas segundo o valor de *Máxima Relevancia* poidan presentar unha gran redundancia, é dicir, a dependencia entre as características pode ser grande. Cando dúas características teñen unha forte dependencia entre elas, o poder do

discriminador de clase non cambia excesivamente se se elimina unha delas. Polo tanto, pódese engadir a condición de *mínima Redundancia (Min-Redundancy)* para seleccionar características mutuamente exclusivas, como se pode ver na ecuación (3.7):

$$\min R(S), \qquad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \tag{3.7}$$

A combinación das dúas restricións anteriores dá lugar ao criterio denominado Minima Redundancia Máxima Relevancia (mRMR) [53]. Defínese o operador $\phi(D, R)$ para combinar D e R e optimizar D e R simultaneamente dun xeito máis sinxelo segundo a ecuación (3.8):

$$\max \phi(D, R), \qquad \phi = D - R \tag{3.8}$$

Na práctica, pódense empregar métodos de busca incrementais para atopar as características case óptimas definidas por ϕ. O algoritmo de mRMR non pretende seleccionar funcións que sexan independentes unhas das outras, sen embargo, en cada paso intenta seleccionar a función que minimice a redundancia e maximice a relevancia. Para datos reais, as características seleccionadas desta forma tería máis ou menos correlación co resto deles.

O método mRMR evita a difícil estimación da densidade multivariante no cálculo da *Dependencia Máxima*. Ademais, este método pode combinarse eficazmente con outros métodos de selección de características, como pode ser o caso dos métodos envolventes, coa fin de buscar un subconxunto de características con menores custos computacionais. Por último, cabe sinalar que este método é unha boa aproximación práctica, xa que é capaz de acadar unha boa precisión de clasificación con baixa complexidade computacional.

3.2 Metodos de discretización

Existen certos algoritmos de filtrado que, para ser empregados, requiren que os datos de entrada sexan discretos. Trátase dunha práctica común o feito de discretizar os datos numéricos previo paso de aplicaren a selección de características, xa que tamén pode facer que a aprendizaxe sexa máis rápida e precisa. A todo isto podemos sumar que a discretización dos datos fai que estos sexan moito máis fáciles de entender, empregar e explicar. A discretización simplifica os cálculos de entropía e da información mutua, xa que nos permite traballar con funcións de masa de probabilidade durante un número reducido de intervalos extraídos directamente do reconto no conxunto de adestramento. Este enfoque é suficiente na práctica e por iso é tan frecuente o seu uso. En particular, neste problema que estamos a tratar, empregaremos métodos de selección de características que usan información mutua para calcular a relevancia das características e que por iso requiren da realización deste proceso de discretización dos datos como paso previo para poder ser empregados.

3.2.1 Procedemento de discretización

Nun problema de aprendizaxe automática, os atributos poden ser numéricos ou nominais. O termo numérico fai referencia a valores enteiros e reais polo que, dado que algúns algoritmos de selección de características non se poden empregar con datos continuos, o procedemento de discretización foi collendo peso a medida que aumentaba o foco de atención que está a recibir este ámbito da aprendizaxe máquina. O procedemento de discretización baséase en agrupar valores continuos nun número de intervalos discretos. Este proceso implica tomar varias decisións, como por exemplo cantos valores continuos deben agruparse nun intervalo, cantos intervalos deben construírse ou onde os puntos de corte deben situarse na escala continua de valores. Tomar estas decisións non é trivial e depende do algoritmo escollido.

No caso deste traballo, tomamos a decisión de empregar o algoritmo de discretización no que se seleccionan intervalos de igual ancho. Trátase dun método non supervisado, global e estático, o cal se converteu nun dos métodos máis empregados debido á súa sinxeleza. Divide a liña de números entre v_{min} e v_{max} en k intervalos de igual ancho, de xeito que os intervalos teñan un ancho $w = (v_{max} - v_{min})/k$ e os puntos de corte estean nas posicións $v_{min} + w, v_{min} + 2w, \ldots, v_{min} + (k-1)w$. Neste caso, k é un parámetro definido polo usuario que tomará o valor de 5.

3.3 Metodos de clasificación

Neste novo apartado preséntase a descrición dos diferentes algoritmos de clasificación que empregaremos neste proxecto, os cales nos permitirán avaliar o rendemento da selección de características ao aplicalo sobre o problema en cuestión.

3.3.1 C4.5

O algoritmo C4.5 foi desenvolvido por Quinlan en 1993 [32], como unha extensión do algoritmo ID3 (Iterative Dicotomiser 3) desenvolvido en 1986 [54], e así, baséase en árbores de decisión. Unha árbore de decisión clasifica un exemplo ou mostra, filtrándoo en orde descendente, ata que atopa unha folla que corresponda á clasificación buscada.

O algoritmo comeza colocando unha característica na raíz da árbore. Para decidir que atributo ten unha maior capacidade de discriminar as mostras de adestramento entre as distintas clases, úsase a relación de ganancias (*gain ratio*), que se define como $IG(X_i, c_j)/H(x_i)$, onde X_i é unha característica, c_j unha clase e $IG(X_i, c_j)$ representa a ganancia de información (*information gain*). H representa a entropía, que xa se discutiu en apartados anteriores. Deste xeito pódese evitar que se beneficien na selección as características con maior número de valores posibles.

O seguinte paso é crear unha rama para cada valor posible do atributo en cuestión. As mostras de adestramento distribúense nos nodos descendentes segundo o valor que teñen para o atributo raíz. A continuación, o proceso repítese seleccionando un novo atributo que se colocará en cada un dos nodos xerados. Unha das melloras para C4.5 sobre ID3 é que manexa atributos discretos e continuos. Para xestionar atributos continuos, C4.5 crea un limiar e dependendo do valor que toma o atributo, o conxunto de mostra divídese.

A medida que se engaden niveis, as hipóteses vólvense tan perfeccionadas que describen moi ben os exemplos empregados na aprendizaxe, pero o erro de clasificación pode aumentar ao avaliar os exemplos. Para evitar o efecto indeseado da equiparación, o algoritmo C4.5 incorpora a poda da árbore de clasificación unha vez creada. Ademais de evitar un exceso de axuste, grazas á poda hai un aforro no tempo de execución.

A variante de poda usada polo algoritmo C4.5 chámase *post prunning*. Consiste en que, unha vez que se xerou a árbore completa, hai que expor a peza a podar para mellorar o rendemento e de paso obter unha árbore máis curta. Para ser aplicado, o C4.5 debe converter a árbore nun conxunto mínimo de regras.

O clasificador C4.5 ten un parámetro chamado *Factor de confianza*, que é o valor que se empregará cando se poda a árbore, é dicir, cando se eliminen as ramas que se consideran que proporcionan pouca ou ningunha ganancia na precisión estatística do modelo. Ao facer este factor menor, conseguirase unha poda máis agresiva, mentres que aumentala obterá unha mellor aproximación ao conxunto de adestramento. Nas distintas probas experimentais que se realizarán neste traballo utilizaranse os valores 0,25 (valor predeterminado na ferramenta Weka [55, 56], e tamén o recomendado polo autor do algoritmo).

3.3.2 Naïve Bayes (NB)

O algoritmo Naïve Bayes [57] usa a regra de Bayes, presentada na ecuación (3.9), para determinar a probabilidade de cada clase para un exemplo dado.

$$P(h|S) = \frac{P(S|h)P(h)}{P(S)} \tag{3.9}$$

Onde:

- $P(h)$ é a probabilidade a priori da hipótese h. É dicir, sería a probabilidade de que a unha mostra se lle asignase unha determinada etiqueta.
- $P(S)$ é a probabilidade de observar o conxunto de adestramento S.
- $P(S|h)$ é a probabilidade de observar o conxunto de adestramento S, nun universo onde se verifica a hipótese h.

- $P(h|S)$ é a probabilidade posterior de h, cando se observou o conxunto de adestramento S.

A aprendizaxe bayesiana pode verse como o proceso de atopar a hipótese máis probable, dado un conxunto de exemplos de adestramamento S e un coñecemento a priori da probabilidade de cada hipótese. A aplicación do teorema de Bayes para a clasificación consiste en calcular a hipótese con maior probabilidade *a posteriori*, como se mostra na ecuación (3.10).

$$h_{MAP} = \arg\max_{h \in H} \frac{P(S|h)P(h)}{P(S)} \tag{3.10}$$

Onde S rerpesenta os datos de adestramento e h as hipóteses, ou neste caso cada unha das etiquetas a asignar ás mostras. O subíndice MAP significa o máximo a posteriori. Deste xeito, a instancia clasifícase como a que ten maior probabilidade a posteriori.

Dise que o modelo *naive Bayes* é sinxelo ou ”inxenuo” porque supón que os atributos son condicionalmente independentes uns dos outros, dada a clase. Así, as probabilidades a posteriori pódense calcular seguindo a aproximación da ecuación (3.11).

$$P(x_1, x_2, ..., x_n|c_j) = \prod_i P(x_i|c_i), \tag{3.11}$$

Sendo c_j cada unha das clases a clasificar e x_i os valores dos datos cos que se clasifica.

As vantaxes dos clasificadores naive Bayes son que son sinxelos, eficientes e robustos fronte ao ruído e atributos irrelevantes. Ademais, precisan unha pequena cantidade de datos de entrada para estimar os parámetros necesarios para a clasificación.

Os principais inconvenientes deste método son a necesidade dun coñecemento a priori e o elevado custo computacional. A pesar de ser un método cunha restrición tan forte como a independencia dos atributos, pódense obter clasificadores precisos incluso cando non se cumpre esa condición.

Este clasificador ten un parámetro que pode ser variado, é o parámetro do núcleo, que se se establece en true, significa que un estimador do núcleo usarase para atributos numéricos en vez dunha distribución normal.

3.3.3 IB1

O algoritmo IB1 [58] forma parte da familia de algoritmos chamados IBL (*Instance Based Learning*) que se caracteriza por expresar a aprendizaxe como o propio conxunto de exemplos de adestramento. Os algoritmos da familia IBL almacenan exemplos na memoria como puntos no espazo dimensional n, definidos polos n atributos que describen os exemplos, e nunca cambian a representación deses puntos. As dúas decisións máis relevantes tomadas polos

algoritmos desta familia son: Que puntos hai que almacenar? e, que métrica adoptar para medir a semellanza entre os exemplos?

Todas as variantes da familia IBL usan a técnica "veciño máis próximo" para clasificar novos exemplos. Este enfoque de aprendizaxe pode considerarse como unha extensión do algoritmo NN (*Nearest Neighbor*, veciño máis próximo) [59].

O algoritmo IB1 [60] é practicamente idéntico ao algoritmo NN, xa que se limita a buscar o caso almacenado máis próximo (xeralmente segundo unha determinada métrica de distancia euclidiana, ecuación (3.12)) ao exemplo para ser clasificado.

$$d_E(I_1, I_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.12}$$

sendo $I1$ e $I2$ dous exemplos dun espazo bidimensional, de coordenadas $(x1, y1)$ e $(x2, y2)$ respectivamente.

A nova mostra asígnase á clase da instancia de mostra recuperada. Este algoritmo usa métodos eficientes de emparellamento para recuperar estes casos almacenados para que poidan aplicarse en novas situacións. No que respecta ao algoritmo NN tamén posúe as seguintes características:

- Normaliza o dominio dos atributos.
- Procesa de xeito incremental as mostras.
- Ten unha política de tolerancia aos valores de atributos que faltan.
- Almacena todas as mostras de adestramento, que se procesan de xeito incremental.

Unha descrición da aprendizaxe baseada en mostras componse do conxunto de mostras de adestramento almacenadas e, eventualmente, dalgunha información relativa ao desempeño previo das mostras durante o proceso de clasificación. Ese conxunto de mostras pode cambiar unha vez procesada cada mostra de adestramento. Cada instancia está representada por un conxunto de pares atributo-valor e unha clase asociada.

En principio, todas as mostras están descritas polo mesmo conxunto de n atributos e cada mostra x_i é un vector n-dimensional. Na descrición de aprendizaxe, unha clase componse do conxunto de todas as mostras que teñen o mesmo valor para o atributo de clase. Os algoritmos IBL asumen que mostras similares teñen clasificacións similares. Isto implica o uso dunha heurística local para a clasificación do veciño 'máis semellante'.

Do mesmo xeito que o algoritmo Naïve Bayes, a aprendizaxe baseada na semellanza é xeralmente sinxela computacionalmente e as variacións son moitas veces consideradas como modelos de aprendizaxe humana [61].

3.3.4 Support Vector Machine (SVM)

O algoritmo SVM (*Support Vector Machine*) é unha técnica de clasificación [62, 63] que se basea na idea de minimización de riscos estruturais (SRM, *Structural Risk Minimization*) [64]. En moitas aplicacións, o uso do algoritmo SVM mostrou un gran rendemento, máis que as técnicas de aprendizaxe tradicionais como as redes neuronais [62] e introducíronse como ferramentas poderosas para resolver problemas de clasificación.

O algoritmo SVM aprende a superficie de decisión de dúas clases diferentes de puntos de entrada. Coa descrición dada polos datos dos vectores de soporte, é capaz de formar unha fronteira de decisión arredor do dominio dos datos de aprendizaxe sen apenas ou ningún coñecemento dos datos fóra desta fronteira. Primeiro, os datos son mapeados mediante un *kernel Gaussiano* ou outro tipo de kernel a un espazo de características nun espazo de maior dimensión (é dicir, se os puntos de entrada están en R^2, entón son mapeados por SVM a R^3), onde se busca a máxima separación entre clases. Atópase un hiperplano que os separa e maximiza a marxe m entre as clases deste espazo como se mostra na Figura 3.2. Esta función de fronteira, devolta ao espazo de entrada, pode separar os datos en todas as diferentes clases, cada unha formando un agrupamento.

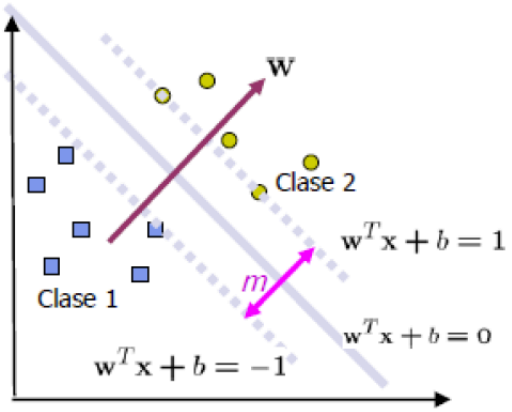

Figura 3.2: A fronteira de decisión dun SVM debe estar tan lonxe dos datos de ambas clases como sexa posible.

Maximizar a marxe m é un problema de programación cuadrática (*QP, Quadratic Programming*) e pode resolverse mediante o seu problema dual introducindo multiplicadores de Lagrange. A solución do hiperplano óptimo pode escribirse como a combinación duns poucos puntos de entrada chamados *vectores de soporte*.

3.3.5 Random Forest (RF)

O algoritmo de Random Forest [65], como o seu nome indica, consta dun gran número de árbores de decisión individual que funcionan como un conxunto. Cada árbore individual no Random Forest esconde unha predición de clase e a clase con máis votos convértese na predición do noso modelo. Nestas árbores contaremos con dous tipos de aleatoriedade. En primeiro lugar, cada árbore está baseada nunha mostra aleatoria dos datos orixinais. En segundo lugar, en cada nodo de árbore, un subconxunto de características selecciónase ao azar para xerar a mellor división. Esta aleatoriedade permite a creación de varios modelos non relacionados, característica que constitúe a clave do funcionamento do algoritmo. Isto é, un gran número de modelos (árbores) relativamente non correlacionados que funcionan como comité, superando calquera dos modelos constitutivos individuais.

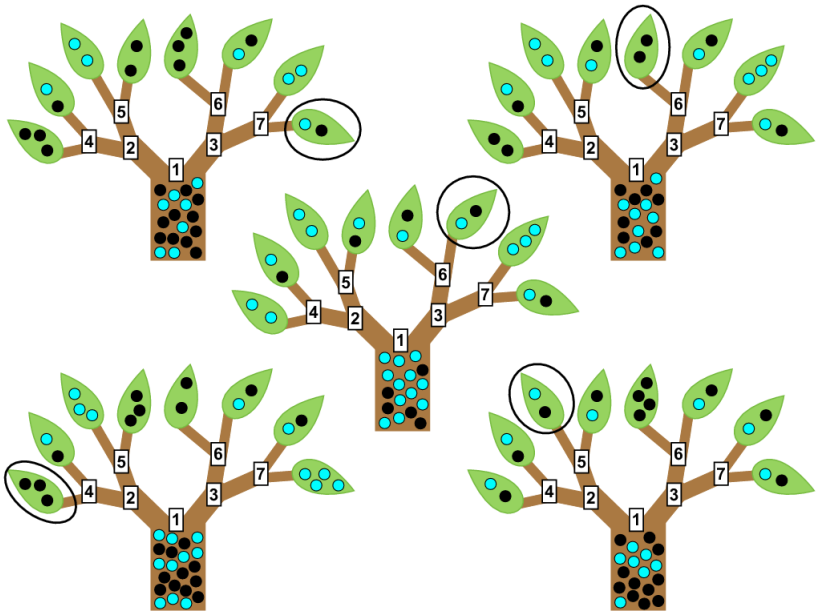

Figura 3.3: Exemplo dun Random Forest con cinco árbores de decisión. Cada árbore foi adestrada nun conxunto de datos de obxectos negros ou azuis. Nótese que o conxunto de adestramento para cada árbore é diferente dos demais. Supoñamos que queremos clasificar un obxecto como negro ou azul e que acaba na folla que está rodeada. Entón a probabilidade de que este obxecto sexa negro é a porcentaxe de obxectos negros en todas as follas (8/11)

Capítulo 4

Experimentación e resultados

Neste capítulo falaremos dos pasos pertencentes á metodoloxía que seguimos para realizar os diferentes experimentos plantexados neste proxecto, e analizaranse os resultados obtidos nos mesmos.

4.1 Metodoloxía experimental

Ao longo das seccións deste apartado explicaranse en detalle cada un dos pasos principais da metodoloxía seguida e do proceso de experimentación levado a cabo. A Figura 4.1 mostra graficamente a metodoloxía experimental seguida ao longo deste traballo.

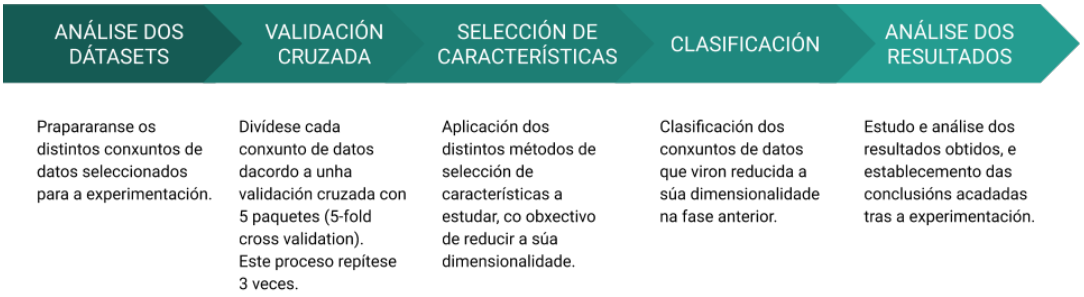

Figura 4.1: Metodoloxía seguida na experimentación.

4.1.1 Preparación do conxunto de datos

O rendemento das distintas configuracións propostas está probado nunha ampla mostra composta por 56 conxuntos de datos, sendo sete deles datasets microarrays, que se consideran

representativos dos problemas que poden xurdir no mundo real. A descrición detallada destes conxuntos de datos pódese atopar nas Táboas 2.1 e 2.2.

Cada un destes conxuntos divídese segundo a validación cruzada de k-fold. Esta técnica proponse, en comparación cunha división conxunta de adestramento e proba, para poder facer uso de todos os datos dispoñibles, pero mantendo en cada iteración unha parte dos datos para probar os modelos xerados para a partición de adestramento. Neste caso, establécese o valor $k = 5$ xa que é un dos valores máis empregados na literatura. Polo tanto, con esta técnica, o conxunto de datos divídese en cinco particións mutuamente excluíntes, que conteñen aproximadamente o mesmo número de mostras. Realízanse cinco validacións e en cada unha delas queda un dos subconxuntos para a parte de test, e o sistema adéstrase cos catro restantes. Así, a precisión estimada é a media das cinco precisións obtidas. Pódense obter máis detalles sobre este tipo de validación na Sección 2.3.

4.1.2 Aplicación de métodos de selección de características

Os diferentes experimentos realizados consisten en facer comparacións entre a aplicación dos métodos de selección de características individualmente, así como a selección de características aleatoria, que será a liña base para as nosas comparacións. Para iso, tómanse os resultados obtidos na aplicación dos métodos individuais nos diferentes conxuntos de datos estudados neste traballo. Como se comentou na Sección 3.1.6, elíxense seis métodos de selección de características recoñecidos. Mentres que dous dos métodos de selección de características devolven un subconxunto de características (CFS e INTERACT), os outros catro (IG, ReliefF, mRMR e MIM) son métodos de tipo ranker, é dicir, non seleccionan un subconxunto de características, se non que se establece un ranking ordenado de todas as características do conxunto de datos, polo que é necesario o establecemento dun limiar para obter un subconxunto de características. Neste traballo optamos por manter o 10% e o log2(n) das características máis relevantes do ranking ordenado, onde n é o número de características nun determinado conxunto de datos. No caso dos microarrays, debido ao desaxuste entre a dimensionalidade e o tamaño da mostra, os limiares seleccionados foron o 5% e as características log2(n) superiores, respectivamente. En cada iteración da validación cruzada nos conxuntos de datos, catro das cinco particións úsanse para construír os distintos rankings de características mediante os métodos de selección escollidos. Cada un destes métodos de selección execútase independentemente do resto, de xeito que cada un deles recibe o mesmo grupo de mostras como entrada (as mostras correspondentes ás catro particións de adestramento). Como resultado deste proceso, teremos unha versión reducida en canto a número de características dos diferentes datasets que estamos a empregar, preparados para ser empregados na fase de clasificación que nos permitirá avaliar o noso experimento. Nótese que algúns dos métodos de selección de características que estamos a utilizar só poden traballar sobre datos discretos

e, nese caso, aplicarase unha discretización en 5 intervalos de igual ancho (ver Sección 3.2).

4.1.3 Clasificación dos conxuntos de datos tras a fase de pre-procesado

Por último, tras obter os novos datasets simplificados contituídos so por aquelas características que se consideraron relevantes tras aplicarense os diferentes métodos de selección de características anterioremente descritos, podemos pasar á fase de clasificación na que se farán as probas experimentais que se propuxeron co fin de analizar que configuracións proporcionan os mellores e os peores resultados con respecto ao método aleatorio de selección de caracteríticas. Para isto, empregaremos os métodos de clasificación mencionados no apartado 3.3. Despois do adestramento, o subconxunto de proba, obtido despois de realizar a validación cruzada, úsase para adquirir o valor de precisión medio estimado de clasificación. Este valor emprégase para verificar a adecuación dos experimentos realizados ao longo deste proxecto de fin de grao, e poder realizar así unha comparativa entre os diferentes métodos de selección de características, aplicadas a unha ampla variedade de datasets e sendo clasificados por varios clasificadores distintos.

Neste traballo empregáronse un total de cinco métodos de clasificación para levar a cabo o proceso de avaliación, cada un pertencente a unha familia diferente. Os clasificadores empregados foron: dous lineais (naive Bayes e Support Vector Machine usando un núcleo lineal) e tres non lineais (C4.5, IB1 e Random Forest). Os cinco clasificadores executáronse mediante a ferramenta Weka [55], utilizando valores por defecto para os seus parámetros.

4.2 Resultados experimentais

Nesta sección preséntanse os resultados obtidos tras realizaren as diversas probas experimentais empregando os conxuntos de datos de referencia mostrados na Sección 2.2, levándose a cabo un estudo crítico de ditos resultados. Dito estudo baséase na comparativa das porcentaxes medias estimadas da precisión de test, analizando principalmente a variabilidade nos métodos de selección de características empregados frente ao método aleatorio (representado como 'Ran' nas táboas). A continuación e durante os seguintes apartados preséntanse unha serie de gráficas comparativas co obxetivo de facilitar a visualización dos resultados obtidos. Todos os resultados se poden ver con máis detalle nas táboas presentes no seu correspondente apartado.

Detallaranse os resultados obtidos con cada método de clasificación empregado. Inicialmente realízase unha análise comparativa entre os clasificadores empregados neste traballo, cos diferentes limiares asociados a cada experimento, para coñecer cales aportaron mellores resultados.

4.2.1 Resultados dos datasets normais

A continuación realizaráse unha análise dos resultados sobre os datasets chamados "normais" (ver Táboa 2.1) en función de cada un dos clasificadores empregados na experimentación, onde se inclúen diferentes táboas nas que se mostra para cada dataset empregado a precisión de clasificación en función do método de selección de carácteristicas que foi aplicado na fase de pre-procesado. Por outra banda, e para facer máis sinxela e visual a interpretación que facemos dos resultados, tamén se presenta de forma gráfica o resultado dos test estadísticos levados a cabo nos que se amosa e avalía como de significativamente diferentes son entre si os diferentes métodos de selección de características empregados.

4.2.1.1 Resultados para o clasificador C4.5

Esta sección baséase na análise dos resultados obtidos a partir dos experimentos realizados co clasificador C4.5. Se observamos os datos presentes na Táboa 4.1 podemos ver que dous dos métodos empregados obtiveron a maior precisión practicamente no 50% dos datasets que conforman os experimentos. Para o clasificador C4.5, o método con mellor porcentaxe de acerto foi INTERACT, sendo o mellor en 13 dos 49 datasets empregados, e seguido por CFS que o foi en 11 dos mesmos. Tamén cabe destacar que o método mRMR con limiar logarítmico obtivo o mellor resultado en 9 dos datasets empregados. Con todo isto, todos os métodos empregados obtiveron unha maior precisión que o método aleatorio (Ran) excepto nalgún dos experimentos realizados, nos cales os valores de precisión distaban tan so unhas centésimas.

Sobre os resultados dos experimentos sobre o clasificador C4.5 que estamos a tratar, leváronse a cabo uns test estadísticos para obervar cales dos métodos presentan diferencias significativas entre eles, especialmente en comparación co método aleatorio, que é a nosa liña base. Nestes diagramas de diferencia crítica, os métodos que non presentan diferenzas significativas entre si están agrupados na mesma barra, e son mellores canto máis á dereita se presentan.

Como se pode observar de forma moi clara na Figura 4.2, o método aleatorio, tanto con limiar do 10% como con log2, non presenta diferenzas significativas entre si, mais si que se observa a distancia que gardan con respecto ao resto dos métodos empregados, xa que son significativamente peores. A outra gran diferenza que se pode observar é a distancia que separa aos dous métodos que tiveron a mellor actuación nos experimentos, que foron INT e CFS que se sitúan á dereita da gráfica e sen diferenzas significativas entre eles, pero si cos resto dos métodos que se agrupan na parte central da grafica en cuestión.

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
arcene	72.83	73.17	75.50	69.00	**76.33**	65.33	69.33	63.83	76.00	74.67	76.00	75.17
arrhythmia	**68.58**	66.82	67.85	65.56	68.07	61.57	55.85	54.21	68.51	62.09	68.58	62.98
basehock	90.13	91.14	**91.60**	87.00	84.26	55.73	69.39	52.18	87.25	69.29	87.17	70.25
bc-wisc-diag	94.09	**94.14**	90.28	92.74	92.80	93.38	85.65	90.57	92.45	93.09	93.44	93.09
bc-wisc-prog	74.24	74.24	73.74	73.74	75.41	72.38	**76.26**	74.21	74.25	73.73	74.09	73.73
breast	93.14	**93.43**	92.38	92.91	89.92	92.97	84.53	88.15	92.85	93.03	92.38	92.91
carcinom	70.54	72.24	71.49	67.28	70.53	60.57	62.29	38.50	72.42	67.25	72.61	**75.71**
COIL20	**92.55**	92.11	88.70	76.53	86.00	64.44	92.15	81.18	89.47	84.38	89.51	90.30
congress	95.25	**95.33**	94.79	94.87	95.02	95.02	81.84	87.51	94.87	94.87	**95.33**	94.87
connbenchsonar	74.84	72.76	69.58	69.58	**75.33**	**75.33**	65.53	65.53	70.99	70.99	71.47	71.47
connect4	70.50	**77.44**	68.30	69.34	68.76	69.40	65.90	65.92	68.30	69.34	68.13	69.34
dermatology	94.89	**95.91**	62.39	75.52	70.78	76.27	57.60	69.24	62.39	75.14	75.15	75.14
gisette	93.88	93.67	94.27	90.77	94.56	91.69	90.09	61.29	94.31	90.74	**94.25**	92.20
glass	64.82	**65.91**	50.30	63.58	45.80	65.78	41.11	57.64	45.80	65.76	45.80	65.76
heart	80.86	80.00	72.59	83.46	70.00	78.77	66.54	75.68	75.80	**84.07**	75.80	**84.07**
hillvalley	49.07	49.07	49.07	49.01	49.07	49.01	**49.12**	49.01	49.01	49.01	49.01	49.01
ionosphere	90.69	90.41	88.04	90.70	83.95	89.74	80.43	86.14	88.61	**90.78**	90.59	**90.78**
isolet	**83.21**	79.76	68.42	42.94	61.49	41.64	71.46	37.17	59.91	38.18	60.00	48.68
krvskp	94.09	**96.62**	90.43	90.58	89.81	92.77	60.89	62.34	90.43	90.34	90.43	90.34
landsat	83.94	**84.41**	78.12	78.78	58.18	58.65	72.25	75.96	78.12	78.72	78.07	78.72
libras	**65.65**	64.35	39.63	35.09	34.91	22.87	56.85	53.15	37.50	22.96	37.22	36.76
lowresspectR	**83.55**	82.23	79.60	78.40	78.47	76.52	80.41	78.41	78.59	79.53	79.03	80.91
molecbiolpromoter	77.40	78.04	79.62	79.62	**81.49**	**81.49**	64.72	64.72	80.52	80.52	80.84	80.84
molecbiolsplice	**93.69**	93.59	93.43	93.43	92.03	92.03	60.22	60.22	93.43	93.43	93.43	93.43
musk2	95.75	96.05	**96.23**	95.85	94.83	94.54	95.10	93.67	95.03	93.19	95.07	94.42
nci9	35.00	28.33	35.00	38.33	33.89	37.78	32.22	16.67	47.78	54.44	47.22	**55.00**
optdigits	90.44	**90.46**	75.68	75.68	77.09	77.09	50.58	50.58	76.16	76.16	76.17	76.17
orlraws10P	74.67	75.00	63.33	64.33	70.33	60.67	72.33	67.67	67.33	76.00	68.67	**78.67**
ozone	96.69	96.32	97.08	**97.12**	**97.12**	**97.12**	**97.02**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**
parkinsons	85.98	87.01	85.30	**86.32**	84.44	84.96	77.95	78.80	85.13	85.81	84.27	85.81
page_blocks	**97.05**	96.98	93.65	96.67	90.09	94.21	91.13	94.60	91.30	95.68	91.30	95.69
PCMAC	82.09	85.35	**86.29**	81.95	81.37	56.22	60.64	50.78	76.55	63.92	76.57	68.97
pendigits	96.06	95.96	57.11	76.30	58.32	77.54	49.76	80.39	57.11	76.30	57.11	76.30
pixraw10P	90.33	91.00	91.00	88.67	**92.00**	81.67	88.33	90.67	89.67	89.33	88.00	90.00
RELATHE	81.01	81.94	**82.71**	75.38	77.53	59.05	62.60	55.01	80.78	60.99	80.85	69.12
satimage	**85.92**	85.84	83.30	83.66	60.86	61.45	77.34	80.68	83.30	83.75	83.59	83.80
segmentation	**96.54**	96.42	79.80	86.67	80.06	86.70	66.19	83.58	75.79	81.72	75.79	81.72
semeion	**75.92**	73.01	63.99	43.44	63.55	33.59	62.46	38.04	65.62	43.61	65.60	56.33
sonar	75.94	75.12	74.34	74.34	**76.25**	**76.25**	65.35	65.35	75.78	75.78	76.09	76.09
soybeansmall	97.11	97.11	**100.00**	**100.00**	96.30	96.30	60.89	70.30	**100.00**	**100.00**	**100.00**	**100.00**
spect	80.65	**80.89**	79.40	79.65	79.40	78.65	79.40	79.40	79.40	79.40	79.40	79.40
splice	93.63	**93.74**	93.73	93.73	92.02	92.02	63.11	63.11	93.73	93.73	93.73	93.73
USPS	**88.78**	87.01	67.99	54.91	79.00	69.13	84.53	72.75	81.37	53.23	81.26	78.36
warpAR10P	64.87	63.08	65.64	59.49	**69.23**	57.18	61.79	48.21	72.31	62.05	72.05	**69.23**
warpPIE10P	81.11	81.27	79.68	73.81	79.05	74.60	79.52	60.63	**81.75**	72.54	81.59	81.43
waveform	**77.43**	76.11	67.89	69.17	71.34	74.73	55.26	57.11	68.60	69.17	68.60	69.17
wine	93.66	**94.03**	78.66	92.33	69.23	91.39	51.53	81.45	78.66	92.33	78.66	92.33
yale	43.03	45.05	47.68	43.64	42.22	40.61	42.22	32.32	47.47	50.30	47.68	**48.89**
zoo	91.10	**92.38**	83.81	88.06	65.95	77.16	49.73	72.90	83.81	89.06	83.81	89.06

Táboa 4.1: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador C4.5. Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% ou 10% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

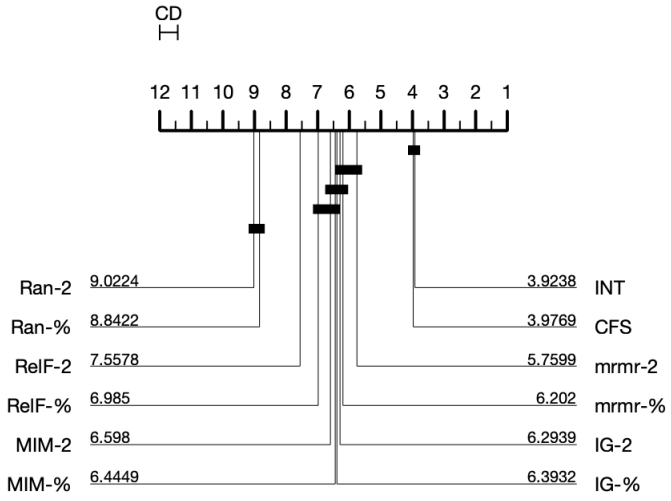

Figura 4.2: Test estatístico para o clasificador C4.5.

4.2.1.2 Resultados para o clasificador Naïve Bayes

Nesta sección comentaranse os resultados obtidos tras os experimentos levados a cabo en base ao clasificador Naïve Bayes. Se observamos a Táboa 4.2 onde se atopan os resultados dos experimentos asociados ao clasificador en cuestión, podemos ver como hai certa semellanza en canto os resultados obtidos co clasificador C4.5, xa que o método aleatorio segue a ser superado en precisión polo resto de métodos empregados na experimentación, destacando ao igual que no caso previo CFS e INT sobre o resto dos métodos, coa diferencia de que nesta ocasión CFS obtivo a maior precisión case no 45% dos datasets empregados nos experimentos. Para obter unha información máis útil e detallada de cara a facer un análise da situación que se comenta, podemos observar a Figura 4.3 pertencente aos test estadísticos realizados sobre os experimentos que estamos a avaliar. Podemos ver que de igual forma que no caso anterior, o método aleatorio segue a obter os peores resultados, situándose máis a esquerda da gráfica e sen mostrar diferenzas significativas entre os limiares empregados na execución. É no caso dos métodos con mellores resultados onde atopamos a principal diferencia con respecto ao caso anterior, xa que neste caso, a pesares de que tanto CFS e INT distan moito en precisión do resto dos métodos empregados, obsérvase que si que existe diferenza significativa entre ambos métodos, sendo CFS o que destaca como o máis preciso na experimentación. Tamén se pode observar como, para todos os métodos de tipo ranking, o limiar do 10% funciona mellor que o logarítmico.

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
arcene	70.83	67.67	67.83	69.33	70.83	57.17	67.00	65.17	68.50	66.00	68.50	**74.17**
arrhythmia	68.21	67.26	65.33	62.61	64.30	58.19	53.98	53.32	**68.73**	61.81	**68.73**	64.53
basehock	89.87	88.12	89.23	81.80	79.08	56.40	69.28	54.31	**90.95**	70.60	**90.95**	71.50
bc-wisc-diag	94.44	94.44	92.21	94.44	93.85	94.61	84.77	89.75	**94.55**	94.14	94.20	94.14
bc-wisc-prog	74.88	74.38	71.06	70.03	72.23	71.03	74.58	73.90	**75.57**	72.21	75.56	72.21
breast	**93.50**	92.09	92.79	92.73	90.92	92.62	84.29	87.75	92.97	92.38	93.14	92.38
carcinom	83.37	82.03	83.96	74.37	81.28	61.38	68.45	46.73	**84.14**	79.90	**84.14**	83.15
COIL20	96.67	**96.69**	79.33	64.81	77.96	51.62	87.75	66.39	85.07	72.20	85.07	85.14
congress	93.33	92.41	92.80	93.18	92.72	**94.18**	80.61	85.44	92.87	93.10	93.18	93.10
connbenchsonar	64.85	65.35	68.25	68.25	**70.14**	**70.14**	62.29	62.29	69.21	69.21	69.21	69.21
connect4	**65.83**	59.15	**65.83**	**65.83**	**65.83**	**65.83**	65.18	65.02	**65.83**	**65.83**	**65.83**	**65.83**
dermatology	**97.30**	97.21	61.92	75.25	69.66	76.56	55.64	67.94	61.92	75.24	74.59	75.24
gisette	**93.17**	88.80	90.18	86.71	88.00	84.31	75.93	57.98	90.25	86.85	90.24	90.41
glass	48.43	**51.54**	47.52	48.75	46.26	45.94	37.03	49.07	46.26	49.53	46.26	49.53
heart	83.58	**84.94**	71.73	83.21	69.14	78.89	65.19	74.57	75.56	83.95	75.56	83.95
hillvalley	48.19	48.19	48.13	47.97	47.91	**48.35**	48.08	47.97	**48.35**	48.19	**48.35**	47.86
ionosphere	**91.74**	88.98	85.76	88.61	82.54	87.18	71.90	73.13	85.47	86.61	88.70	86.61
isolet	**75.28**	67.10	43.76	21.47	35.12	25.88	57.50	25.10	32.36	22.45	32.36	30.68
krvskp	**93.23**	87.33	86.92	86.11	90.43	92.07	59.86	60.95	85.89	85.89	90.43	85.89
landsat	76.61	**77.55**	74.72	73.01	55.33	55.96	68.70	71.69	74.72	73.93	73.47	73.93
libras	**62.87**	60.93	25.65	26.76	23.24	19.63	47.96	42.87	23.15	19.91	23.15	24.17
lowresspectR	**81.61**	80.79	78.15	75.02	75.58	71.18	76.27	75.84	78.40	78.09	78.40	78.78
molecbiolpromoter	**88.40**	88.10	86.84	86.84	86.49	86.49	63.84	63.84	87.78	87.78	87.78	87.78
molecbiolsplice	91.77	**91.83**	90.05	90.05	90.61	90.61	56.22	56.22	90.05	90.05	90.05	90.05
musk2	68.10	82.11	83.87	**86.33**	83.28	81.53	78.33	79.95	81.27	78.91	81.27	82.90
nci9	34.44	32.78	28.33	40.00	28.89	36.67	27.78	20.00	53.89	63.89	53.89	**73.33**
optdigits	**91.56**	90.95	68.57	68.57	71.17	71.17	44.69	44.69	70.40	70.40	70.40	70.40
orlraws10P	90.33	**91.33**	81.33	67.67	83.33	65.67	**91.33**	78.33	90.00	81.67	**91.33**	88.67
ozone	78.31	79.55	84.33	86.57	69.14	72.59	**89.50**	91.22	69.70	73.07	69.70	72.70
parkinsons	77.78	77.61	84.27	82.22	81.88	73.68	67.18	66.67	**85.13**	81.37	81.03	81.37
page_blocks	**94.38**	90.82	90.04	93.85	89.88	89.97	89.69	85.74	90.85	92.04	90.85	92.04
PCMAC	78.64	77.73	78.57	73.73	73.36	56.17	59.84	52.22	**80.58**	62.00	**80.58**	69.62
pendigits	**83.32**	82.90	44.49	59.16	48.56	62.91	43.45	64.99	44.49	59.16	44.49	59.16
pixraw10P	96.00	**97.33**	97.00	90.00	96.67	89.00	**97.33**	94.33	**97.33**	91.33	**97.33**	90.00
RELATHE	81.50	79.21	81.71	63.16	75.10	56.79	61.86	54.61	**83.39**	57.82	**83.39**	69.26
satimage	**79.38**	78.23	76.66	76.39	57.52	57.73	70.98	74.77	76.66	76.42	75.86	76.42
segmentation	**85.25**	80.20	62.40	68.02	67.00	70.84	50.40	63.98	55.80	55.90	55.80	55.90
semeion	**84.93**	79.91	61.77	41.83	62.79	29.34	65.56	35.84	63.23	41.64	63.23	57.12
sonar	67.48	67.66	65.40	65.40	70.34	**70.34**	63.12	63.12	69.07	69.07	69.07	69.07
soybeansmall	**99.33**	98.59	97.85	97.11	96.30	97.78	60.81	70.96	97.85	97.11	99.26	97.11
spect	76.66	75.05	74.40	73.67	74.40	75.78	**79.40**	77.27	76.65	74.16	77.03	74.16
splice	93.64	**93.97**	91.56	91.56	90.68	90.68	59.72	59.72	91.56	91.56	91.56	91.56
USPS	**83.34**	73.30	43.31	35.75	70.47	58.91	69.76	55.42	71.69	40.13	71.69	73.57
warpAR10P	**78.46**	75.38	67.95	59.49	66.41	55.38	63.59	38.72	68.21	57.44	68.21	62.05
warpPIE10P	**93.97**	93.97	85.24	55.87	87.14	70.48	81.75	58.25	92.38	51.27	92.38	91.59
waveform	**78.91**	78.59	66.16	66.10	71.71	74.97	54.83	56.49	67.03	66.10	67.03	66.10
wine	97.56	**97.94**	78.66	94.01	70.74	93.64	51.88	83.13	78.66	94.01	78.66	94.01
yale	55.96	54.75	54.55	45.05	50.30	39.19	55.76	37.78	**56.16**	50.91	**56.16**	51.92
zoo	**95.38**	95.35	87.10	89.10	65.95	76.16	49.70	73.24	87.10	91.71	87.10	91.71

Táboa 4.2: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador Naïve Bayes. Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% ou 10% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

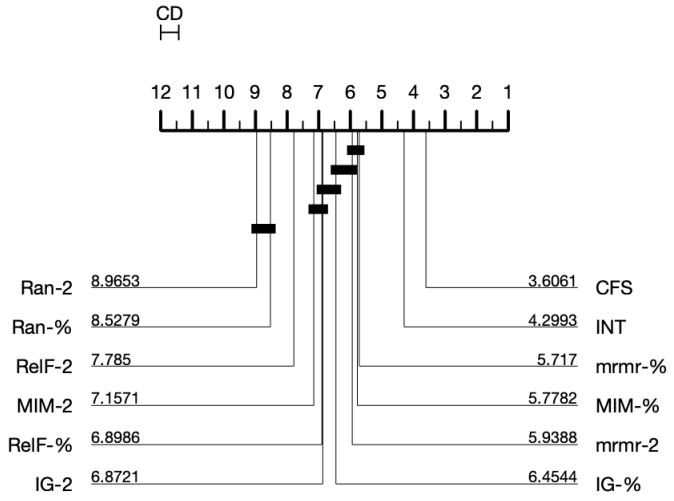

Figura 4.3: Test estatístico para o clasificador Naïve Bayes.

4.2.1.3 Resultados para o clasificador IB1

Nesta sección analizamos os resultados obtidos a partir dos experimentos realizados co clasificador IB1. Se observamos os resultados que obtivemos nesta experimentación podemos ver que o método aleatorio non conseguiu acadar a mellor precisión de clasificación en ningunha das execucións coas diferentes bases de datos empregadas (Táboa 4.3). Ao igual que nos casos anteriores, volven a ser os métodos CFS e INT, nesa orde respectivamente, os que mellores resultados acadan. Neste caso, ao contemplar a Figura 4.4 na que se amosan de forma gráfica as diferenzas significativas entre os diferentes métodos, vemos que, de xeito similar ao que ocorría co clasificador C4.5, o método aleatorio obtén os peores resultados (sen diferenzas significativas entre os dous limiares) e de novo CFS e INT son mellores que o resto dos métodos aínda que non se producen diferenzas significativas entre eles. Se obviamos o método de selección aleatoria, vemos como, entre os métodos analizados, ReliefF presenta os peores resultados, especialmente co limiar logarítmico. Igual que para o clasificador Naive Bayes, para todos os métodos de tipo ranking, o limiar do 10% funciona mellor que o logarítmico.

4.2.1.4 Resultados para o clasificador SVM

Nesta sección analizamos os resultados obtidos tras os experimentos levados a cabo en base ao clasificador SVM. De xeito similar ao que puidemos observar nos clasificadores xa comentados anteriormente, o método aleatorio segue a ser o que peores resultados presenta

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
arcene	81.17	76.00	80.33	78.83	**84.83**	68.50	80.33	67.83	81.50	76.33	81.50	80.67
arrhythmia	64.90	64.09	**67.03**	62.83	60.70	58.04	55.68	52.88	65.12	62.09	65.12	62.09
basehock	**91.89**	89.31	89.75	89.30	76.10	56.43	70.25	54.02	85.85	69.51	85.85	70.16
bc-wisc-diag	**95.55**	95.20	90.33	94.26	93.44	95.31	85.41	90.64	93.38	95.08	93.91	95.08
bc-wisc-prog	72.40	72.73	70.39	72.39	70.89	72.04	67.88	68.71	70.88	73.58	**75.56**	73.58
breast	93.85	93.50	92.15	93.44	90.04	93.20	84.64	88.39	92.85	**94.02**	92.56	**94.02**
carcinom	89.85	89.87	**90.06**	77.99	89.86	64.61	79.52	41.97	89.28	82.76	89.28	84.67
COIL20	99.47	**99.61**	96.50	81.23	90.56	68.15	98.10	83.77	97.48	88.77	97.48	93.84
congress	95.10	93.56	94.79	93.72	95.02	95.56	82.38	87.66	95.02	93.79	**95.63**	93.79
connbenchsonar	81.52	**79.78**	73.08	73.08	77.83	77.83	70.48	70.48	72.12	72.12	72.12	72.12
connect4	70.46	**78.84**	68.31	69.46	68.74	69.32	65.91	65.95	68.31	69.46	68.13	69.46
dermatology	**96.18**	95.72	62.30	75.98	70.31	76.73	56.57	66.64	62.30	76.81	74.59	76.81
gisette	95.61	95.59	96.67	90.84	96.60	91.89	87.03	58.89	**96.79**	90.91	96.79	92.49
glass	**68.08**	65.90	53.26	64.19	47.67	60.93	44.22	55.93	47.67	63.41	47.67	63.41
heart	79.14	78.77	72.35	80.49	69.75	78.02	66.05	73.33	75.56	**81.85**	75.56	**81.85**
hillvalley	50.71	50.71	**52.64**	50.99	50.93	49.00	49.39	49.51	52.14	47.58	52.14	49.83
ionosphere	88.79	85.66	87.66	86.52	84.33	87.18	79.66	83.29	87.66	87.56	**90.21**	87.56
isolet	**60.44**	52.46	55.32	42.33	50.03	40.41	39.10	35.95	47.95	37.51	47.95	48.58
krvskp	94.07	**95.07**	90.43	90.43	89.89	92.77	60.95	62.52	90.43	90.18	90.34	90.18
landsat	85.64	**87.22**	78.20	78.78	57.58	58.03	72.11	75.70	78.20	78.71	78.17	78.71
libras	**75.74**	75.37	40.09	36.39	33.52	21.02	66.39	59.54	35.46	22.50	35.46	37.69
lowresspectR	**87.38**	87.07	81.73	79.85	82.61	77.21	81.67	82.05	80.60	82.48	80.60	81.61
molecbiolpromoter	80.84	80.22	83.71	83.71	82.74	82.74	63.92	63.92	**86.51**	**86.51**	**86.51**	**86.51**
molecbiolsplice	68.82	68.87	88.27	88.27	87.61	87.61	55.67	55.67	**88.27**	**88.27**	**88.27**	**88.27**
musk2	94.76	95.30	**95.50**	95.22	94.15	93.60	94.31	92.50	94.49	93.32	94.49	93.59
nci9	47.78	50.00	51.11	43.33	56.67	42.78	40.56	20.00	**66.11**	52.78	**66.11**	61.11
optdigits	**98.62**	97.66	75.00	75.00	77.92	77.92	49.05	49.05	76.90	76.90	76.90	76.90
orlraws10P	95.00	95.00	91.33	75.67	95.00	70.33	91.67	73.00	**95.67**	84.33	95.67	93.00
ozone	96.41	96.42	96.75	**96.79**	96.52	96.64	96.77	96.70	96.62	96.56	96.62	96.66
parkinsons	85.98	**87.86**	85.13	86.15	84.44	86.32	78.29	79.49	85.13	87.35	86.32	87.35
page_blocks	95.62	**95.77**	93.76	95.61	89.89	93.89	90.90	93.78	91.05	94.45	91.05	94.45
PCMAC	**86.98**	80.87	83.77	83.63	65.81	56.72	60.87	51.53	80.70	65.05	80.70	69.94
pendigits	**99.14**	99.10	56.25	76.43	56.63	76.62	46.85	80.78	56.25	76.43	56.25	76.43
pixraw10P	**99.00**	**99.00**	90.33	87.00	95.00	86.33	96.33	87.67	91.33	87.67	91.33	85.33
RELATHE	82.15	77.86	81.27	75.12	77.20	59.40	65.43	56.23	**83.04**	60.92	**83.04**	69.33
satimage	**90.26**	88.27	82.93	84.26	58.90	59.78	77.27	81.97	82.93	84.39	83.15	84.38
segmentation	95.58	**95.66**	80.71	87.55	78.76	87.11	64.13	82.84	79.49	85.43	79.49	85.43
semeion	**87.99**	82.07	64.24	43.63	65.37	33.27	67.13	36.93	66.25	43.94	66.25	56.69
sonar	77.38	**78.36**	74.82	74.82	75.95	75.95	65.85	65.85	72.74	72.74	72.74	72.74
soybeansmall	99.26	97.85	**100.00**	95.70	93.41	96.37	57.48	67.48	**100.00**	97.78	97.78	97.78
spect	80.28	**81.66**	79.40	79.65	77.89	78.15	79.40	79.65	79.40	78.53	79.40	78.53
splice	74.33	74.35	**89.10**	**89.10**	88.01	88.01	60.23	60.23	**89.10**	**89.10**	**89.10**	**89.10**
USPS	**95.96**	92.71	67.63	56.53	85.25	73.77	89.71	67.37	85.91	52.70	85.91	79.83
warpAR10P	73.85	71.03	75.38	70.00	**77.44**	64.10	47.69	34.87	74.62	70.26	74.62	69.23
warpPIE10P	**97.78**	**97.78**	95.08	78.25	94.60	83.65	92.38	72.54	95.56	78.57	95.56	92.70
waveform	**81.33**	79.23	67.77	68.67	71.41	74.53	54.38	55.27	68.57	68.67	68.57	68.67
wine	94.76	**94.94**	78.66	94.21	69.60	93.28	52.67	82.94	78.66	94.21	78.66	94.21
yale	55.15	56.36	56.16	53.33	50.51	43.64	47.68	33.54	59.60	55.76	59.60	**60.20**
zoo	**90.46**	89.75	83.81	83.14	65.29	74.83	50.06	67.94	83.81	85.43	83.81	85.43

Táboa 4.3: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador IB1. Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% ou 10% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

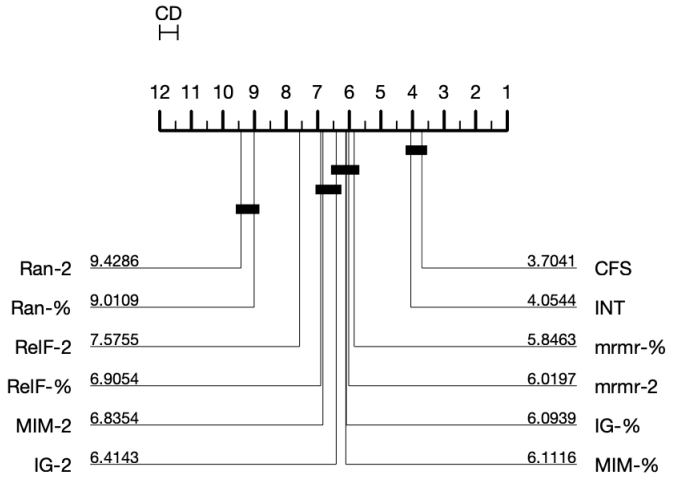

Figura 4.4: Test estatístico para o clasificador IB1.

en canto a precisión de clasificación con respecto ao resto de métodos empregados nos experimentos (Táboa 4.4). Observando en detalle a Figura 4.5, podemos ver tres grandes grupos. O primeiro grupo está formado por CFS e INT que son, unha vez máis, os que presentan os mellores resultados, neste caso cunha importante diferenza cos demais. O segundo grupo está comprendido polo resto dos métodos, nos que vemos como as barras se solapan, aínda que métodos como mRMR e MIM funcionan mellor que ReliefF co limiar logarítmico. E, por último, e como era de esperar, de novo a selección aleatoria é a que obtén os peores resultados, con diferenza significativa respecto ao resto de filtros.

4.2.1.5 Resultados para o clasificador Random Forest

Nesta sección comentaranse os resultados obtidos tras os experimentos levados a cabo co clasificador Random Forest. Como podemos observar na Táboa 4.5 de resultados, INT e CFS seguen a ser os métodos que maior precisión alcanzan nos experimentos, vendo que, tal e como indican os test estadísticos (Figura 4.6), non se observan diferenzas significativas entre eles, mais si que presentan unha ampla diferenza con respecto ao resto dos métodos empregados. Pola outra banda, Ran2 (selección aleatoria con limiar algorítmico) continúa a ser o método que menor precisión presenta nos resultados, tendo neste caso unha diferenza significativa con respecto á súa versión co limiar do 10%, que nesta ocasión non presenta diferenzas con respecto ao método ReliefF con limiar logarítmico.

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
arcene	77.00	72.33	83.50	68.00	79.67	57.83	80.17	62.67	**83.33**	67.83	83.17	75.33
arrhythmia	68.15	65.41	67.70	59.66	**68.51**	62.39	58.92	55.38	68.22	60.03	68.29	59.88
basehock	91.37	91.27	**94.28**	81.77	84.58	56.53	76.00	51.52	92.69	70.30	92.69	70.38
bc-wisc-diag	**96.78**	96.49	91.15	94.38	93.68	95.67	84.01	90.75	94.20	94.09	94.38	94.09
bc-wisc-prog	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**	**76.26**
breast	**94.55**	94.38	92.79	92.56	89.63	93.67	83.82	88.45	92.97	92.50	92.91	92.50
carcinom	93.52	94.08	92.54	75.50	93.11	60.78	89.29	46.97	**95.42**	80.12	95.42	79.73
COIL20	**99.98**	99.95	94.68	59.12	86.83	49.31	98.84	68.91	96.04	67.52	96.04	82.73
congress	94.18	94.18	93.72	94.10	93.72	94.33	80.61	84.98	93.72	93.95	**94.48**	93.72
connbenchsonar	76.24	75.94	72.57	72.57	**76.25**	**76.25**	63.11	63.11	73.85	73.85	73.69	73.69
connect4	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**	**65.83**
dermatology	96.65	**96.84**	61.18	76.07	71.42	77.28	53.96	66.45	61.18	76.72	75.98	76.72
gisette	95.47	93.17	**96.55**	88.21	96.04	87.10	91.49	60.76	96.45	88.21	96.45	90.77
glass	**50.38**	50.37	37.87	46.77	35.53	48.79	30.55	40.04	35.53	48.32	35.53	48.32
heart	84.57	83.21	72.22	84.32	69.14	78.89	65.19	74.07	75.56	85.56	75.56	**84.94**
hillvalley	47.36	47.36	51.81	50.60	**53.19**	52.75	51.43	50.05	51.43	48.57	49.94	48.40
ionosphere	**87.94**	87.84	82.34	86.70	76.66	79.02	71.14	76.46	82.15	83.38	81.59	83.38
isolet	**85.31**	74.78	59.57	23.02	57.54	33.02	66.92	23.63	53.99	26.69	53.99	35.55
krvskp	94.06	**94.08**	90.43	90.67	89.92	91.91	60.31	61.03	90.43	90.43	90.43	90.43
landsat	83.11	**84.67**	77.11	77.34	56.80	56.86	68.97	72.55	77.10	77.50	76.00	77.39
libras	**64.81**	60.74	22.50	20.28	20.46	13.24	42.04	33.70	20.09	13.33	20.28	20.37
lowresspectR	**87.63**	86.63	82.36	76.96	80.03	71.81	79.36	75.02	82.61	81.10	82.49	82.49
molecbiolpromoter	81.77	81.46	82.42	82.42	83.00	83.00	60.81	60.81	**83.97**	**83.97**	**83.97**	**83.97**
molecbiolsplice	84.35	**84.45**	81.16	81.16	82.86	82.86	53.65	53.65	81.04	81.04	81.47	81.47
musk2	87.89	87.73	**90.98**	87.22	88.65	84.89	86.91	84.59	84.59	84.59	84.59	84.59
nci9	50.56	50.00	52.78	41.11	57.22	42.78	51.11	17.78	**72.78**	63.89	**72.78**	57.22
optdigits	**97.89**	96.00	74.39	74.39	74.69	74.69	48.19	48.19	74.69	74.69	74.66	74.66
orlraws10P	98.67	**99.67**	95.33	38.00	96.00	33.67	98.00	70.00	98.67	50.33	98.67	86.00
ozone	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**	**97.12**
parkinsons	84.96	**87.35**	82.22	85.64	81.03	82.74	75.04	75.73	82.39	83.42	81.54	83.25
page_blocks	92.23	**92.66**	89.79	92.09	89.83	91.44	90.06	90.77	89.93	92.22	89.93	92.27
PCMAC	79.02	82.47	**86.10**	70.61	74.88	55.94	64.28	51.16	84.78	63.87	84.78	67.92
pendigits	96.89	**96.92**	49.47	67.59	53.54	71.98	44.17	70.69	49.47	67.59	49.47	67.59
pixraw10P	**99.00**	98.67	97.00	32.67	96.00	47.33	**99.00**	83.67	91.67	34.67	91.67	35.00
RELATHE	83.60	81.71	83.53	58.98	78.16	57.88	69.07	56.27	85.89	61.01	**85.94**	67.13
satimage	**85.82**	84.63	82.55	82.73	58.15	58.34	75.04	79.32	82.55	82.78	82.15	82.77
segmentation	**90.14**	89.94	55.86	67.27	61.93	70.66	48.33	63.41	50.56	53.62	50.56	53.62
semeion	**88.51**	82.53	66.81	44.05	67.84	32.12	67.98	37.22	68.17	43.56	68.24	58.32
sonar	**78.01**	75.79	71.62	71.62	75.76	75.76	67.28	67.28	71.31	71.31	71.48	71.48
soybeansmall	**97.85**	**97.85**	85.78	91.41	96.30	97.78	58.74	66.67	85.78	97.78	97.78	97.78
spect	82.28	**82.78**	79.40	79.40	77.89	77.90	79.40	79.40	79.40	79.40	79.40	79.40
splice	86.94	**87.55**	86.61	86.61	86.85	86.85	58.17	58.17	86.50	86.50	86.31	86.31
USPS	**94.91**	92.37	57.70	44.57	83.76	69.99	89.00	67.20	85.31	47.77	85.30	79.33
warpAR10P	**87.44**	84.62	87.69	53.59	86.41	44.62	91.79	37.44	83.33	48.72	83.85	57.69
warpPIE10P	98.57	**98.73**	96.51	60.63	96.35	70.95	**98.73**	59.21	96.51	52.70	96.51	89.52
waveform	**84.21**	84.11	68.41	70.57	71.58	75.53	55.43	57.39	69.42	70.56	69.29	70.57
wine	95.87	**96.06**	78.66	93.26	68.83	93.26	50.70	81.84	78.66	93.26	78.66	93.26
yale	64.65	57.58	65.66	37.78	58.59	34.95	61.62	28.69	**65.25**	36.97	64.24	43.03
zoo	91.71	**92.06**	62.94	75.79	65.95	76.16	48.16	66.00	62.94	74.48	62.94	74.48

Táboa 4.4: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador SVM. Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% ou 10% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
arcene	81.17	80.17	79.33	74.00	**82.33**	69.83	80.00	70.17	81.50	78.50	81.50	**82.33**
arrhythmia	**73.82**	73.08	71.90	68.95	72.79	61.80	63.80	52.82	72.42	66.00	73.01	67.18
basehock	94.68	94.61	**96.99**	89.66	89.05	56.67	78.96	54.44	94.18	69.74	94.10	70.23
bc-wisc-diag	**96.19**	95.96	90.33	94.08	92.62	95.20	87.17	92.51	92.97	93.97	93.79	94.26
bc-wisc-prog	70.86	71.03	72.74	75.57	70.71	74.07	71.04	73.21	71.89	74.74	75.56	**75.58**
breast	93.73	**94.02**	92.27	93.32	90.04	93.03	84.58	88.74	92.79	93.96	93.20	93.96
carcinom	87.96	87.57	87.78	78.17	88.15	63.63	77.43	48.87	87.57	82.20	**88.53**	85.83
COIL20	99.86	**99.91**	98.94	86.04	93.56	70.83	99.28	89.00	99.03	92.94	99.10	96.71
congress	95.33	**95.71**	95.02	94.94	95.40	95.17	82.07	87.82	95.10	95.02	95.48	95.02
connbenchsonar	76.42	**77.21**	72.76	72.76	74.33	74.33	70.19	70.19	71.80	71.80	72.76	72.76
connect4	70.48	**81.37**	68.30	69.46	68.74	69.32	65.91	65.94	68.30	69.46	68.13	69.46
dermatology	**96.28**	96.19	62.11	74.77	70.22	76.55	56.29	68.40	62.11	74.67	74.40	74.67
gisette	95.95	96.11	96.83	91.28	96.69	92.39	93.52	58.79	96.90	91.22	**96.84**	92.93
glass	70.89	**69.80**	45.64	63.26	41.75	65.13	40.17	60.30	41.75	68.56	41.75	68.56
heart	79.01	80.12	72.35	80.25	69.75	78.52	66.05	74.44	75.56	**81.23**	75.56	80.86
hillvalley	51.87	51.87	51.21	**52.47**	49.51	49.01	50.49	48.51	50.99	48.30	50.82	48.68
ionosphere	93.26	93.26	87.94	90.04	84.24	89.75	81.66	88.13	88.13	90.79	**91.64**	90.79
isolet	**93.60**	91.10	79.34	48.07	68.52	44.57	87.28	47.69	66.65	42.49	66.60	52.84
krvskp	94.15	**96.59**	90.43	90.45	89.81	92.77	61.24	62.91	90.43	90.20	90.34	90.20
landsat	86.36	**88.18**	78.32	78.89	57.91	58.19	72.31	75.89	78.31	78.87	78.21	78.86
libras	**77.96**	76.76	44.35	42.22	41.39	29.81	69.91	64.63	44.17	30.83	45.56	44.91
lowresspectR	**88.38**	87.88	82.99	81.29	83.68	79.54	84.00	83.06	82.61	82.55	83.61	83.17
molecbiolpromoter	85.51	86.75	84.94	84.94	87.43	87.43	70.43	70.43	90.91	90.91	**92.16**	**92.16**
molecbiolsplice	94.89	**95.05**	93.29	93.29	91.43	91.43	61.30	61.30	93.29	93.29	93.32	93.32
musk2	96.64	97.06	**97.17**	95.84	96.01	94.99	96.55	94.89	96.48	94.77	96.54	95.63
nci9	47.78	46.11	47.22	42.22	55.56	43.33	40.56	18.89	67.78	65.00	65.00	**74.44**
optdigits	**97.95**	97.28	78.23	78.23	80.65	80.65	50.71	50.71	79.69	79.69	79.63	79.63
orlraws10P	96.67	**97.67**	92.33	74.00	91.33	66.67	96.67	86.67	95.33	83.67	**97.67**	93.00
ozone	96.98	**97.02**	96.95	96.74	96.87	96.79	97.00	96.94	96.90	96.91	96.92	96.94
parkinsons	85.98	**89.74**	85.13	84.79	84.27	85.13	77.61	80.34	85.13	85.64	86.15	85.81
page_blocks	**97.44**	**97.43**	93.70	96.94	89.10	94.52	90.54	94.85	90.22	96.12	90.22	96.13
PCMAC	88.82	89.66	**91.58**	84.20	83.89	57.13	66.82	50.90	86.82	65.16	86.93	70.00
pendigits	**99.01**	98.98	55.55	76.62	56.62	77.63	45.74	82.25	55.55	76.62	55.55	76.62
pixraw10P	**99.67**	**99.67**	96.00	91.33	97.00	89.67	99.00	96.67	98.00	94.00	98.33	91.00
RELATHE	85.54	86.43	**87.57**	74.84	82.85	59.78	72.34	56.25	85.68	60.87	85.68	69.33
satimage	**91.46**	90.13	82.71	84.19	59.56	60.79	77.96	83.09	82.71	84.27	83.21	84.30
segmentation	**98.14**	97.88	82.94	88.96	80.84	89.11	69.65	87.20	80.22	86.88	80.22	86.88
semeion	**90.44**	85.00	67.67	44.36	68.34	33.69	72.65	37.60	69.24	44.57	69.45	56.31
sonar	**79.80**	79.16	71.13	71.13	74.65	74.65	65.84	65.84	70.00	70.00	69.52	69.52
soybeansmall	98.52	98.52	**100.00**	**100.00**	96.30	96.30	61.56	73.04	**100.00**	**100.00**	**100.00**	**100.00**
spect	80.02	**80.16**	79.40	79.41	77.89	77.90	79.40	79.03	79.40	78.53	79.40	78.53
splice	95.40	**95.41**	93.60	93.60	91.46	91.46	63.37	63.37	93.57	93.57	93.53	93.53
USPS	**95.35**	94.11	75.68	62.44	86.36	76.97	92.30	80.55	87.97	60.19	87.97	84.62
warpAR10P	80.51	79.74	80.51	70.51	**81.03**	64.36	75.64	57.44	78.97	67.44	82.56	78.21
warpPIE10P	97.46	**98.41**	95.40	83.02	94.92	88.25	97.62	81.90	95.24	80.00	95.71	90.63
waveform	81.39	**81.55**	67.78	68.55	71.31	74.13	53.81	54.68	68.63	68.52	68.62	68.51
wine	**97.77**	97.39	78.66	94.39	69.60	92.33	52.66	82.75	78.66	94.39	78.66	94.57
yale	**68.28**	66.26	65.66	53.33	61.21	49.09	67.68	48.69	65.86	60.20	67.07	63.64
zoo	95.37	**94.70**	85.76	89.71	65.95	77.16	50.70	70.22	85.76	90.38	85.76	90.38

Táboa 4.5: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador Random Forest. Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% ou 10% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

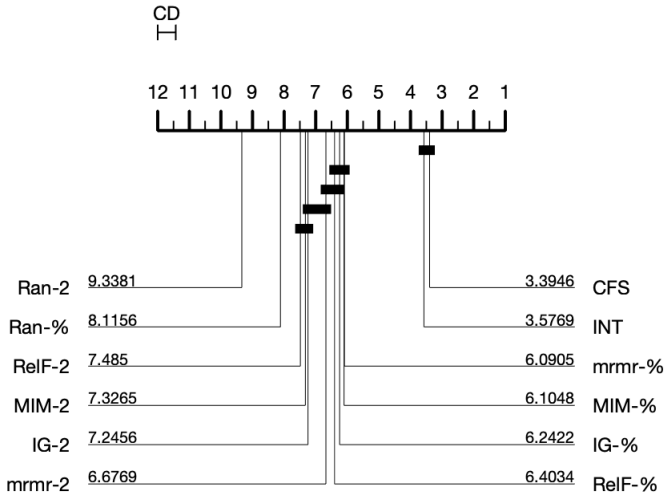

Figura 4.5: Test estatístico para o clasificador SVM.

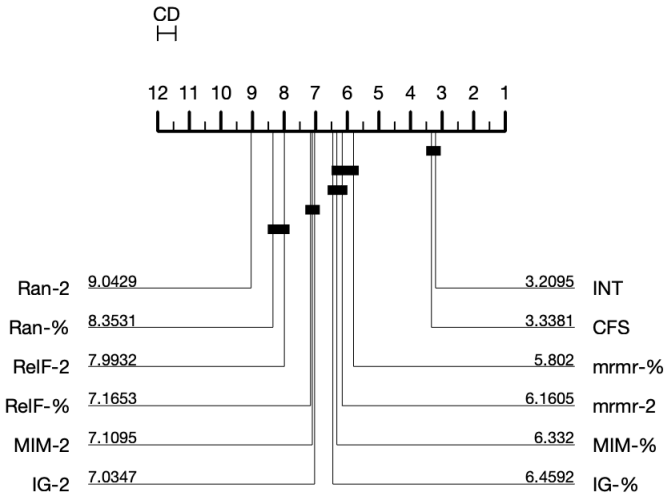

Figura 4.6: Test estatístico para o clasificador Random Forest.

4.2.1.6 Características seleccionadas

Na Táboa 4.6 podemos observar o número de características seleccionadas en media para os datasets considerados "normais". Con estes datos vese reflexada a problemática de usar un

limiar axeitado para os métodos de tipo ranker, xa que en datasets con moitas características (como por exemplo *arcene*), o limiar de porcentaxe selecciona un número de características excesivo, mentras que o limiar logarítmico selecciona un número demasiado baixo, e o máis axeitado parece o seleccionado polos métodos de subconxunto (CFS e INTERACT). Algo similar pasa cos datasets que teñen moi poucas características (como por exemplo *bc-wisc-diag*), para os que o n úmero de características seleccionados polos limiares, tanto de porcentaxe coma logarítmico, son moito menores que o número de características que seleccionan os métodos de subconxunto. É posible que o bo comportamento dos métodos CFS e INTERACT sobre estes conxuntos se deba á súa capacidade de establecer automaticamente o número de características óptimo a seleccionar, e non dependan da necesidade de establecer un limiar axeitado, que é un problema que aínda non está resolto na literatura especializada sen incurrir en procesos moi custosos de comprobar a precisión de clasificación para cada subconxunto de características resultante de ir engadindo cada característica do ranking.

	CFS	INT	Ranker percent.	Ranker log2
arcene	73.53	36.93	500	13
arrhythmia	24.80	21.87	28	8
basehock	58.73	77.93	243	12
bc-wisc-diag	10.47	11.00	3	5
bc-wisc-prog	1.73	1.87	3	5
breast	8.07	11.27	3	5
carcinom	312.93	253.87	459	13
COIL20	183.47	128.27	102	10
congress	3.93	9.67	2	4
connbenchsonar	16.33	13.13	6	6
connect4	6.33	37.40	4	5
dermatology	16.53	15.60	3	5
gisette	77.27	49.13	250	12
glass	5.20	5.40	1	3
heart	6.53	9.27	1	4
hillvalley	1.00	1.00	10	7
ionosphere	9.33	14.40	3	5
isolet	186	57.07	62	9
krvskp	6.33	17.33	4	5
landsat	20.80	33.07	4	5
libras	26.87	17.87	9	6
lowresspectR	18.60	15.73	10	7
molecbiolpromoter	7.47	7.93	6	6
molecbiolsplice	27.53	28.00	6	6
musk2	13.73	18.67	17	7
nci9	45.67	33.93	486	13
optdigits	36.73	21.60	6	6
orlraws10P	415.20	249.47	515	13
ozone	18.40	18.67	7	6
parkinsons	6.47	8.47	2	4
page_blocks	5.93	9.60	1	3
PCMAC	51.80	89.60	164	12
pendigits	13.00	13.53	2	4
pixraw10P	415.53	244.07	500	13
RELATHE	76.73	75.33	216	12
satimage	24.07	12.53	4	5
segmentation	7.67	8.67	2	4
semeion	85.53	39.00	26	8
sonar	11.67	12.67	6	6
soybeansmall	8.87	5.93	4	5
spect	11.40	13.13	2	4
splice	29.87	29.87	6	6
USPS	86.47	32.80	26	8
warpAR10P	47.80	23.67	120	11
warpPIE10P	71.80	52.20	121	11
waveform	14.27	17.60	4	5
wine	10.20	9.67	1	4
yale	39.73	19.60	51	10
zoo	9.20	7.33	2	4

Táboa 4.6: Características seleccionadas para os conxuntos de datos normais

4.2.2 Resultados Microarrays

A continuación realizarase unha análise dos resultados en base a cada un dos clasificadores utilizados na experimentación na que, nesta ocasión, se empregaron conxuntos de datos de tipo microarray para estudar así o comportamento dos distintos métodos de selección de características (en comparación coa liña base que é a selección aleatoria), observando para cada un dos conxuntos de datos empregados a precisión de clasificación en función do método de selección de carácteristicas que foi aplicado na fase de pre-procesado. Para facer máis sinxela e visual a interpretación que facemos dos resultados, realizamos a análise dos test estadisticos levados a cabo acompañados das correspondentes gráficas de xeito que podemos avaliar como de significativamente diferentes son entre si os diferentes métodos de selección de características utilizados na experimentación.

4.2.2.1 Resultados para o clasificador C4.5

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
9_Tumors	31.11	31.11	26.11	29.44	26.11	26.11	23.89	13.89	33.89	42.78	36.11	**45.00**
CNS	58.33	57.22	57.78	58.89	55.00	55.56	58.89	61.67	56.11	73.33	56.11	**76.11**
colon	79.66	79.19	79.66	77.52	77.95	77.86	67.65	56.71	78.93	**83.12**	78.46	78.85
DLBCL	75.70	75.19	75.70	74.44	76.44	79.41	78.22	57.93	77.19	**82.22**	76.44	81.56
Leukemia_1	90.16	89.21	88.76	84.57	86.00	91.56	79.56	58.60	90.22	91.52	90.22	**92.00**
SRBCT	84.36	83.87	84.88	81.74	87.75	82.30	73.70	50.25	89.31	**91.62**	88.53	88.80
TOX_111	59.48	61.22	56.91	51.31	58.13	52.43	55.18	39.42	59.10	54.16	58.93	**61.24**

Táboa 4.7: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador C4.5 (microarrays). Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

Nesta sección analizaremos os resultados obtidos polo clasificador C4.5. Como se pode apreciar na Táboa 4.7, son dous os métodos empregados que concentran a totalidade dos mellores resultados obtidos sobre cada un dos datasets usados. Esta situación pode ser debida a que mRMR foi un método creado específicamente para tratar con datos de tipo microarray e MIM é da mesma familia de métodos (ambos os dous baseados na Teoría da Información). Podemos observar que dos dous limiares utilizados é o logarítimico o que obtén mellor precisión en ambos os métodos a pesares de seleccionar un número de características moito menor que o escollido polo método cando se emprega o limiar do 5% das características. Se observamos os resultados que nos aportan os tests estadísticos (Figura 4.7) podemos ver que a selección aleatoria (Ran2) segue a ser o método que peor precisión obtén e distanciándose considerablemente do resto de métodos que neste caso presentan uns resultados sen diferenzas significativas entre si (na barra máis á esquerda incluso podemos observar que a selección aleatoria con limiar do 5% non presenta diferenzas significativas con outros métodos como Information Gain ou ReliefF).

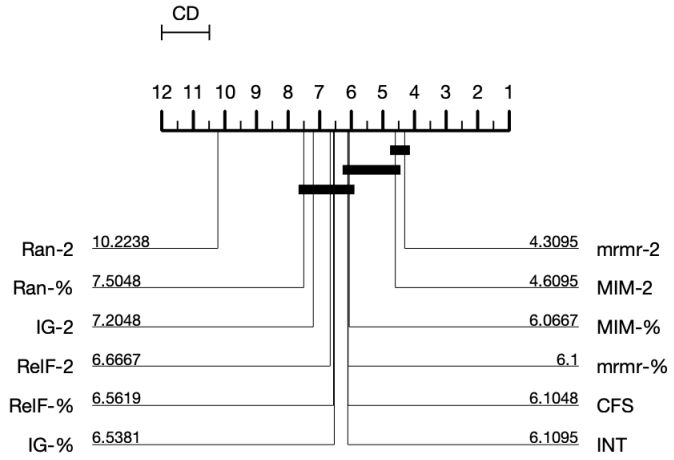

Figura 4.7: Test estatístico para o clasificador C4.5 (microarrays).

4.2.2.2 Resultados para o clasificador Naïve Bayes

Nesta sección analizaremos os resultados obtidos a partir dos experimentos realizados co clasificador Naïve Bayes. Se observamos os datos presentes na Táboa 4.8 e contemplamos a distribución que tomaron os diferentes métodos nos test estadísticos (Figura 4.8), podemos ver como Ran2 e Ran (selección aleatoria) seguen a ser os que, con gran diferenza, son superados polo resto de métodos empregados, presentando os peores valores de precisión na clasificación. De xeito similar ao acontecido no clasificador comentado anterioremente, seguen a ser os métodos MIM e mRMR os que mellores resultados amosan, ainda que as diferenzas significativas existentes con respecto ao resto de métodos se reducen nesta ocasión.

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
9_Tumors	36.67	36.11	40.56	31.11	40.00	34.44	27.22	15.56	48.33	**48.89**	48.33	46.11
CNS	67.78	67.22	57.78	67.78	61.67	61.67	57.78	52.22	68.89	73.89	68.89	**79.44**
colon	79.66	75.98	73.80	80.13	80.17	83.25	58.85	57.74	81.79	**86.45**	81.79	85.98
DLBCL	94.15	94.96	97.04	91.56	97.04	92.89	86.44	61.63	**99.26**	93.63	**99.26**	97.85
Leukemia_1	93.49	93.52	96.70	93.40	96.22	95.30	81.46	58.57	**98.13**	95.27	97.17	97.17
SRBCT	98.01	94.39	98.36	94.75	99.17	94.78	80.44	51.84	**100.00**	94.04	**100.00**	95.25
TOX_111	**79.93**	74.52	66.30	58.67	68.85	61.03	72.55	47.22	68.83	65.69	68.83	73.52

Táboa 4.8: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador Naïve Bayes (microarrays). Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

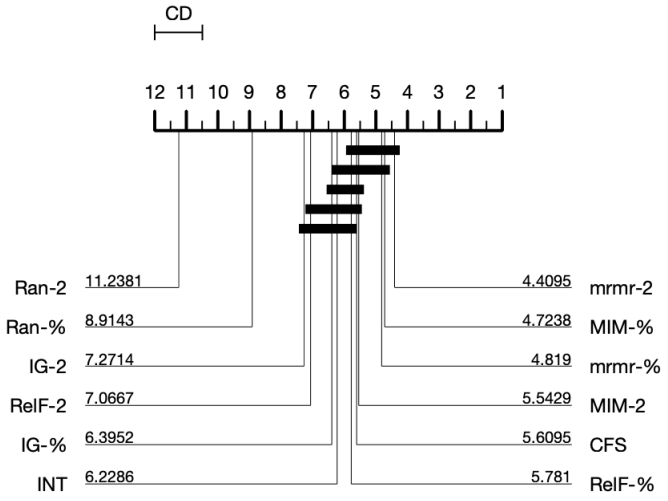

Figura 4.8: Test estatístico para o clasificador Naïve Bayes (microarrays).

4.2.2.3 Resultados para o clasificador IB1

Neste apartado avalíanse os resultados que nos proporcionaron os experimentos levados a cabo empregando o clasificador IB1. Mantendo a dinámica dos casos anteriores, Ran e Ran2 continúan sen superar en precisión a ningún outro dos métodos empregados (Táboa 4.9), mais nesta ocasión Ran non presenta diferenzas significativas con IG2 nos test estadisiticos (Figura 4.9). Pola outra banda, os métodos orientados a microarrays como son MIM e mRMR seguen a concentrar a totalidade dos mellores resultados en canto a precisión no proceso de clasificación.

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
9_Tumors	38.33	38.89	47.22	36.67	46.67	30.00	30.56	18.33	**59.44**	52.22	**59.44**	53.33
CNS	63.89	63.89	61.11	60.56	63.33	65.56	60.56	60.00	**77.22**	74.44	**77.22**	68.89
colon	83.93	78.97	83.21	83.29	79.96	79.96	73.25	66.03	83.76	**84.32**	83.76	82.39
DLBCL	95.63	92.00	94.81	90.67	86.15	91.41	72.22	64.52	97.78	95.78	97.78	**98.52**
Leukemia_1	92.06	90.22	95.33	90.06	93.49	94.32	74.86	51.52	96.22	95.78	96.22	**97.62**
SRBCT	97.62	96.37	99.19	92.35	99.61	96.79	75.07	48.24	**100.00**	96.05	**100.00**	96.47
TOX_111	**88.54**	83.85	77.03	59.24	82.70	61.97	73.65	47.60	79.18	67.27	79.18	71.19

Táboa 4.9: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador IB1. Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

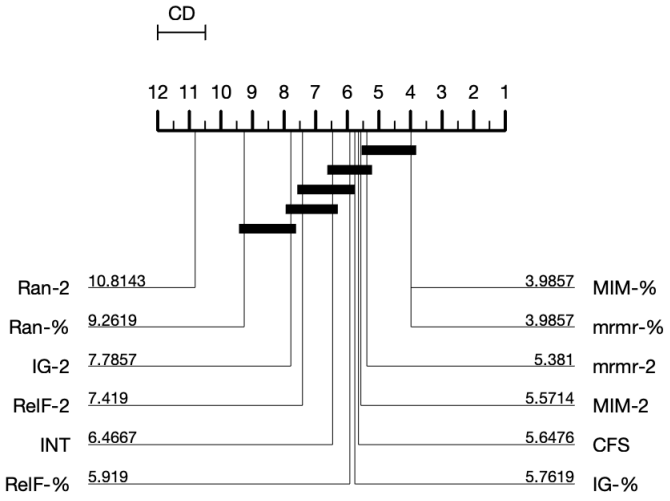

Figura 4.9: Test estatístico para o clasificador IB1 (microarrays).

4.2.2.4 Resultados para o clasificador SVM

Nesta sección avalíanse os resultados obtidos na execución levada a cabo tomando como clasificador a estudar SVM. Ao analizar os datos que se motran na Táboa 4.10 apreciamos como a precisión coa que se clasificou cada dataset supera ás obtidas co resto de clasificadores analizados ata o momento, alcanzando o 100% de precisión para varios dos métodos de selección e en difrentes datasets. Isto pode ser debido a que o SVM demostrou ser un clasificador moi eficaz para tratar con conxuntos de tipo microarray [66]. De novo, Ran e Ran2 volveron a ser as peores opcións, presentando este ultimo grandes diferenzas nos test estadísticos que o separan considerablemente do resto de métodos empregados (Figura 4.10). Nesta ocasión, a pesares de que os métodos MIM e mRMR continuaron obtendo moi bos resultados na súa varíante cun limiar do 5% das mostras, as súas versións logarítmicas foron superadas por métodos como InfoGain ou CFS.

4.2.2.5 Resultados para o clasificador Random Forest

Esta sección baséase na análise dos resultados obtidos a partir dos experimentos realizados co clasificador Random Forest. Se botamos un ollo aos resultados obtidos na experimentación (Táboa 4.11) e a distribución que reflexan os test estadísticos (Figura 4.11), podemos observar como Ran2 segue a ser o método que peores valores de precisión obtivo, seguido de Ran, non presentando este último diferenzas significativas con métodos como IG2 e Rel2. Pola outra

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
9_Tumors	44.44	42.78	58.89	30.00	55.00	30.00	46.67	13.89	**73.33**	43.89	**73.33**	52.78
CNS	66.11	66.11	61.11	61.67	68.89	65.00	63.33	65.00	**84.44**	73.89	**84.44**	67.22
colon	82.78	78.46	83.21	83.76	82.61	83.80	77.69	62.22	87.05	87.01	87.05	**87.56**
DLBCL	95.63	93.48	97.04	92.22	93.48	92.22	86.81	64.37	**100.00**	95.78	**100.00**	**100.00**
Leukemia_1	93.90	92.57	95.30	85.97	94.35	91.59	91.08	62.32	95.75	96.70	95.75	**97.62**
SRBCT	98.77	98.80	99.58	91.99	99.61	98.41	89.61	50.49	**100.00**	95.66	**100.00**	97.23
TOX_111	88.53	85.02	**90.65**	57.69	92.24	58.10	86.78	43.50	90.47	62.37	90.47	75.27

Táboa 4.10: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador SVM (microarrays). Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

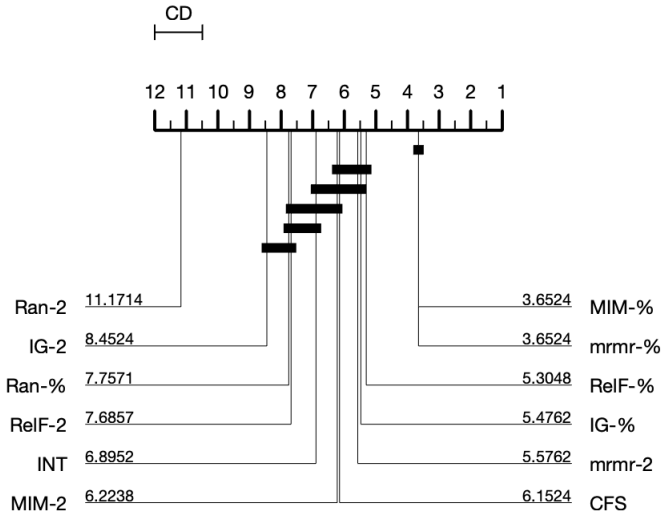

Figura 4.10: Test estatístico para o clasificador SVM (microarrays).

banda, tanto mRMR como MIM continúan a obter os mellores resultados, pero sen chegar a ter diferenzas significativas con métodos como CFS ou Rel nesta ocasión.

4.2.2.6 Características seleccionadas

Na táboa 4.12 podemos ver as características seleccionadas polos distintos métodos avaliados sobre os conxuntos de datos microarray. Por unha banda, para os métodos de tipo subconxunto (CFS e INTERACT), amósase a media para todas as execucións (3 repeticións e 5-fold CV). E, para os métodos de tipo ranker este número é fixo segundo o número de características de cada dataset e amósase nas últimas dúas columnas da táboa.

O limiar logarítmico selecciona un número de características moi reducido, sendo quizais

	CFS	INT	IG-%	IG-2	RelF-%	RelF-2	Ran-%	Ran-2	MIM-%	MIM-2	mrmr-%	mrmr-2
9_Tumors	41.11	40.00	52.22	36.67	48.33	37.22	33.33	18.33	61.67	52.22	**63.89**	52.22
CNS	64.44	63.33	62.78	65.00	61.67	61.11	64.44	60.56	76.11	**78.89**	77.78	78.33
colon	82.78	75.90	80.51	80.56	83.25	**84.87**	71.41	68.72	82.69	84.40	83.76	83.85
DLBCL	91.41	89.26	93.56	86.44	94.22	89.48	88.67	66.67	95.78	93.63	**97.11**	95.04
Leukemia_1	**96.29**	96.25	95.30	94.86	95.27	95.27	81.02	63.37	96.22	95.75	96.25	96.22
SRBCT	97.60	97.18	99.61	92.35	99.22	97.62	89.53	59.19	**100.00**	95.66	99.61	97.62
TOX_111	**79.74**	77.02	74.85	62.40	75.85	63.55	69.64	49.76	75.84	65.89	76.42	72.14

Táboa 4.11: Precisión de clasificación en test (media das 3 repeticións e 5-fold CV) para o clasificador Random Forest (microarrays). Para os métodos de selección que precisan dun limiar, a opción de quedarse co 5% indícase con '-%', e a opción de usar o log2 indícase con '-2'.

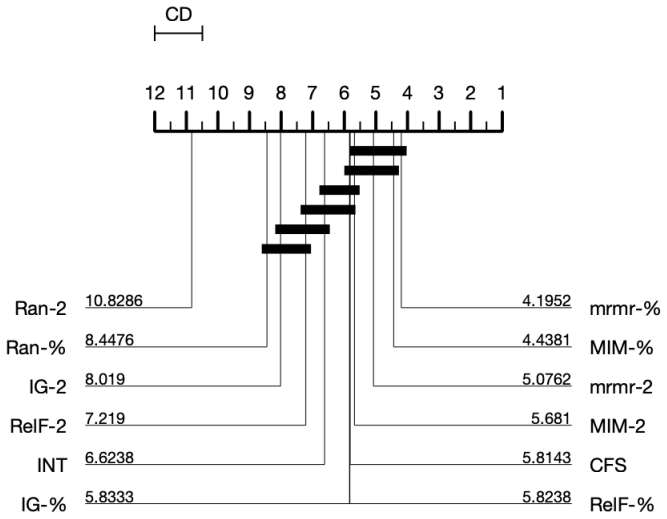

Figura 4.11: Test estatístico para o clasificador Random Forest (microarrays).

a explicación de que nos tests estatísticos case sempre apareza a versión do limiar do 5% cunha precisión superior á acadada polo limiar logarítmico. O número de características seleccionado polos métodos de subconxunto está nun punto medio entre as características do limiar logarítmico e de porcentaxe, como así tamén aparecen as súas precisións en posicións intermedias dos tests estatísticos. A selección dun limiar axeitado para os métodos de tipo ranker é un reto para os investigadores que aínda non está resolto, e que precisa investigación futura.

	CFS	INT	Ranker percent.	Ranker log2
9_Tumors	48.67	37.20	286	12
CNS	42.33	32.27	356	13
colon	23.73	15.13	100	11
DLBCL	64.27	48.27	201	12
Leukemia_1	82.73	44.67	266	12
SRBCT	105.53	65.00	115	11
TOX_111	117.87	76.80	287	12

Táboa 4.12: Características seleccionadas para os conxuntos de datos microarray

Capítulo 5

Conclusións

Neste derradeiro capítulo presentaranse as diferentes conclusións e liñas de traballo futuro ás que se chegaron coa realización deste traballo.

5.1 Conclusións

Este traballo fin de grao parte do obxectivo de estudar de forma exhaustiva os métodos máis populares no campo da selección de características, realizando as pertinentes comparacións entre os mesmos, e comprobando se os resultados obtidos ao empregar o método de selección aleatoria melloran aos proporcionados por algún dos métodos a estudar. Mediante esta experimentación o que se busca é comprobar se realmente se necesitan tantos métodos de selección de características e cales son realmente útiles dependendo dos conxuntos de datos cos que se vaia tratar.

Tras realizar todos os experimentos propostos, mediante os que avaliar as diferentes ideas e hipóteses sobre as que se basea este traballo, podemos confirmar que os métodos analizados son, en practicamente todos os casos, mellores significativamente que a selección aleatoria, confirmando a súa efectividade. Pola outra banda, para os dous tipos de datasets analizados hai métodos que demostran ser claramente superiores (CFS e INTERACT no caso dos datasets normais, e mRMR e MIM no caso dos microarrays).

Canto ao uso do limiar, a grandes rasgos o limiar logarítmico parece máis axeitado para os datasets normais (por detrás dos métodos de subconxunto, que son a opción gañadora para este tipo de datasets) e o limiar de porcentaxe para os datasets microarrays, aínda que neste último dominio non adoita haber diferenzas significativas entre ambos os limiares.

Por outra banda, e dacordo ao teorema No-Free-Lunch, o mellor clasificador non sempre é o mesmo para todos os datasets, polo que tamén se levou a cabo a observación de como se comportaban os diferentes clasificadores en función dos métodos de selección de características empregados na fase de pre-procesado.

Unha vez realizados os experimentos podemos concluir que, tal e como se discute na publicación [67] os resultados que se obteñen na clasificación non foron considerablemente diferentes entre os distintos métodos empregados, mais poderíamos concluir que Random Forest no caso dos datasets normais e SVM no caso dos microarrays foron os que obtiveron, de forma xeral sobre a totalidade dos datasets empregados, os mellores números en canto a precisión de clasificación, ao igual que concluiu Fernández-Delgado no seu estudo [68].

En resumo, e despois de analizar os experimentos realizados, sacamos as seguintes conclusións e podemos facer as seguintes recomendacións cando un usuario desexe executar un proceso de selección de características:

- Demóstrase a efectividade dos métodos de selección de características analizados, xa que sempre obteñen mellores resultados en termos de precisión de clasificación ca facer unha selección aleatoria.

- Se se está a tratar con datasets de tipo "normal", os métodos máis axeitados parecen ser CFS e INTERACT, coa ventaxa adicional de que non hai que establecer un limiar para o número de características a seleccionar.

- Se tratamos con datasets de tipo microarray (datasets con moitas máis características ca mostras e que adoitan ter moitas características redundantes), o máis axeitado é usar os métodos mRMR ou MIM, ambos os dous pertencentes á toolbox FEAST [69]. Estes métodos precisan establecer un limiar para determinar con cantas características se hai que quedar, e parece máis apropiado o limiar de porcentaxe.

- En completa ignorancia das particularidades do problema a resolver, suxerimos o uso de CFS, xa que ademáis de ser un dos dous mellores para datasets de tipo normal, tamén adoitou ser unha boa opción para os microarrays (por detrás de MIM e mRMR), coa vantaxa engadida de non ter que establecer un limiar.

5.2 Traballo futuro

A partir das conclusións obtidas, propóñense as seguintes liñas de investigación como traballo futuro:

- O estudio dun limiar axeitado para os métodos de tipo ranker é un problema importante no campo da selección de características e que aínda está lonxe de resolver. Unha liña de traballo futuro neste eido sería probar un maior número de limiares e o desenvolvemento dun limiar automático para cada tipo de dataset.

- Este estudo experimental pode ser ampliado usando máis datasets de tipo microarray, para ver se as conclusións obtidas se manteñen, e tamén sería interesante usar conxuntos de datos sintéticos, onde se coñecen a priori as características que deben ser seleccionadas e non se precisa o uso de clasificadores para avaliar a efectividade da selección de características.
- Además do estudio da efectividade dos métodos de selección de características en seleccionar as características relevantes dun problema dado, tamén é un traballo futuro interesante o estudo da eficacia destes métodos, interpretada coma a complexidade computacional de cada método de selección de características, e que pode resultar en facelo máis ou menos axeitado para un tipo determinado de datos.

Bibliografía

[1] Broad Institute, "Cancer Program Data Sets," [Online; Febreiro 2020]. Dispoñible: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

[2] Arizona State University, "Feature Selection Datasets," [Online; Febreiro 2020]. Dispoñible:: http://featureselection.asu.edu/datasets.php.

[3] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[4] R. Ruiz Sánchez, "Heurísticas de selección de atributos para datos de gran dimensionalidad," 2006.

[5] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[6] P. Nemenyi, "Distribution-free multiple comparisons (doctoral dissertation, princeton university, 1963)," *Dissertation Abstracts International*, vol. 25, no. 2, p. 1233, 1963.

[7] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[8] R. Bellman, "Dynamic programming," *Princeton University Press*, 1957.

[9] R. E. Bellman, "Adaptive control processes," *Princeton University Press*, 1961.

[10] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based features subset selection: the harder you try the worse it gets," *Proceedings of AI 2004, the 24th SGAI International Conference on Innovate Techniques and Applications of Artificial Intelligence*, 2004.

[11] R. Ruiz-Sánchez, "Heurísticas de selección de atributos para datos de gran dimensionalidad," Ph.D. dissertation, Universidad de Sevilla, Departamento de Lenguajes y Sistemas Informáticos, 2006.

[12] P. Langley, "Selection of redundant features in machine learning," *In Proceedings of the AAAI Fall Symposium on Relevance*, 1994.

[13] N. Wyse, R. Dubes, and A. Jain, "A critical evaluation of intrinsic dimensionality algorithms," in *Pattern Recognition in Practice*, I. E. Gelsema and K. L.N., Eds. Morgan Kaufmann Publishers, 1980, pp. 415–425.

[14] G. Fernández-Bayón, "Selección de variables en sistemas de aprendizaje automático de preferencias," Ph.D. dissertation, Universidad de Oviedo, 2006.

[15] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction. Foundations and Applications.* Springer, 2006.

[16] Z. Zhao and H. Liu, "Searching for interacting features," *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'07)*, pp. 1156–1161, 2007.

[17] A. Blum and P. Langley, "Selection of relevant feature and examples in machine learning," *Artificial Intelligence*, vol. 97(1-2), pp. 245–271, 1997.

[18] G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," *Machine Learning: Proceedings of the 11th International Conference*, pp. 121–129, 1994.

[19] R. Kohavi and G. John, "Wrapper for feature subset selection," *Artificial Intelligence Journal, special issue on relevance*, vol. 97, no. 1-2, pp. 273–324, 1997.

[20] L. Shpigelman, A. Ñavot, N. Tishby, and E. Vaadia, "Nearest neighbor based feature selection for regression and its application to neural activity," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, Cambridge, MA, 2006, pp. 995–1002.

[21] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Occam's razor," *Information Processing Letters*, vol. 24, pp. 377–380, 1987.

[22] D. Gamberger and N. Lavrac, "Conditions for Occam's Razor Applicability and Noise Elimination," *Lecture Notes In Computer Science*, vol. 1224, pp. 108–123, 1997.

[23] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Waikato University, Department of Computer Science, 1998.

[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[25] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems*, pp. 44–49, 1998.

[26] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features." in *AAAI*, vol. 91. Citeseer, 1991, pp. 547–552.

[27] J. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial Intelligence*, vol. 40, pp. 11–61, 1989.

[28] R. Caruana and D. Freitag, "How useful is relevance?" *In Working notes of the AAAI fall symp. on relevance*, pp. 25–29, 1994.

[29] A. Jakulin, "Attribute interactions in machine learning," Ph.D. dissertation, University of Ljubljana, 2003.

[30] A. Jakulin and I. Bratko, "Analyzing attribute dependencies," in *Proc. of Principles of Knowledge Discovery in Data (PKDD)*, ser. LNAI, N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski, Eds., vol. 2838. Springer-Verlag, Sep. 2003, pp. 229–240.

[31] I. Rish, "An empirical study of the naive Bayes classifier," *International Joint Conference on Artificial Intelligence (IJCAI 2001) Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[32] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[33] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning, artificial intelligence97 (1997), 245-271," *Google Scholar Google Scholar Digital Library Digital Library*.

[34] S. J. Russell and P. Norvig, *Inteligencia Artificial: un enfoque moderno*, 2004, no. 04; Q335, R8y 2004.

[35] J. Pearl, "Intelligent search strategies for computer problem solving," *Addision Wesley*, 1984.

[36] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial intelligence*, vol. 151, no. 1-2, pp. 155–176, 2003.

[37] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[38] J. Huang, Y. Cai, and X. Xu, "A wrapper for feature selection based on mutual information," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 2. IEEE, 2006, pp. 618–621.

[39] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[40] J. Pearl, *Heuristics: Intelligent Search Strategies for computer problem solving*. Addison-Wesley, 1984.

[41] Russell and Norvig, *Inteligencia Artificial, un enfoque moderno*. Pearson, 2004.

[42] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," 1998.

[43] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *European conference on machine learning*. Springer, 1994, pp. 171–182.

[44] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 249–256.

[45] C. Shannon, "A mathematical theory of communication." *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[46] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[47] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of 17th International Conference on Machine Learning, ICML2000*, 2000, pp. 856–863.

[48] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1988.

[49] K. Kira and L. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Conference on Machine Learning*, 1992, pp. 249–256.

[50] K. Kira, L. A. Rendell *et al.*, "The feature selection problem: Traditional methods and a new algorithm," in *Aaai*, vol. 2, 1992, pp. 129–134.

[51] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. D. Jong, S. Dzeroski, S. E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. V. D. Welde, W. Wenzel, J. Wnek, and J. Zhang, "The monk's problems: A performance comparison of different learning algorithms," Tech. Rep., 1991.

[52] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[53] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[54] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[55] S. R. Garner *et al.*, "Weka: The waikato environment for knowledge analysis," in *Proceedings of the New Zealand computer science research students conference*, vol. 1995, 1995, pp. 57–64.

[56] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with java implementations," *Acm Sigmod Record*, vol. 31, no. 1, pp. 76–77, 2002.

[57] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[58] L. P. V. Braga, L. I. O. Valencia, and S. S. R. Carvajal, *Introducción a la Minería de Datos*. Editora E-papers, 2009.

[59] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[60] D. Kibler, D. W. Aha, and M. K. Albert, "Instance-based prediction of real-valued attributes," *Computational Intelligence*, vol. 5, no. 2, pp. 51–57, 1989.

[61] S. J. Cunningham, J. Littin, and I. H. Witten, "Applications of machine learning in information retrieval," 1997.

[62] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[63] B. Schölkopf, C. J. Burges, A. J. Smola *et al.*, *Advances in kernel methods: support vector learning*. MIT press, 1999.

[64] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[65] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[66] F. Gonzalez-Navarro, "Feature selection in cancer research: Microarray gene expression and in vivo 1h-mrs domains," *Software department Technical University of Catalonia, Barcelona, Spain*, 2011.

[67] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos, "Do we need hundreds of classifiers or a good feature selection?" in *ESANN*, in press.

[68] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[69] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The journal of machine learning research*, vol. 13, no. 1, pp. 27–66, 2012.