

ESTUDIO EXPERIMENTAL SOBRE ALGORITMOS DE CLASIFICACIÓN SUPERVISADA BASADOS EN CUBRIMIENTO SECUENCIAL

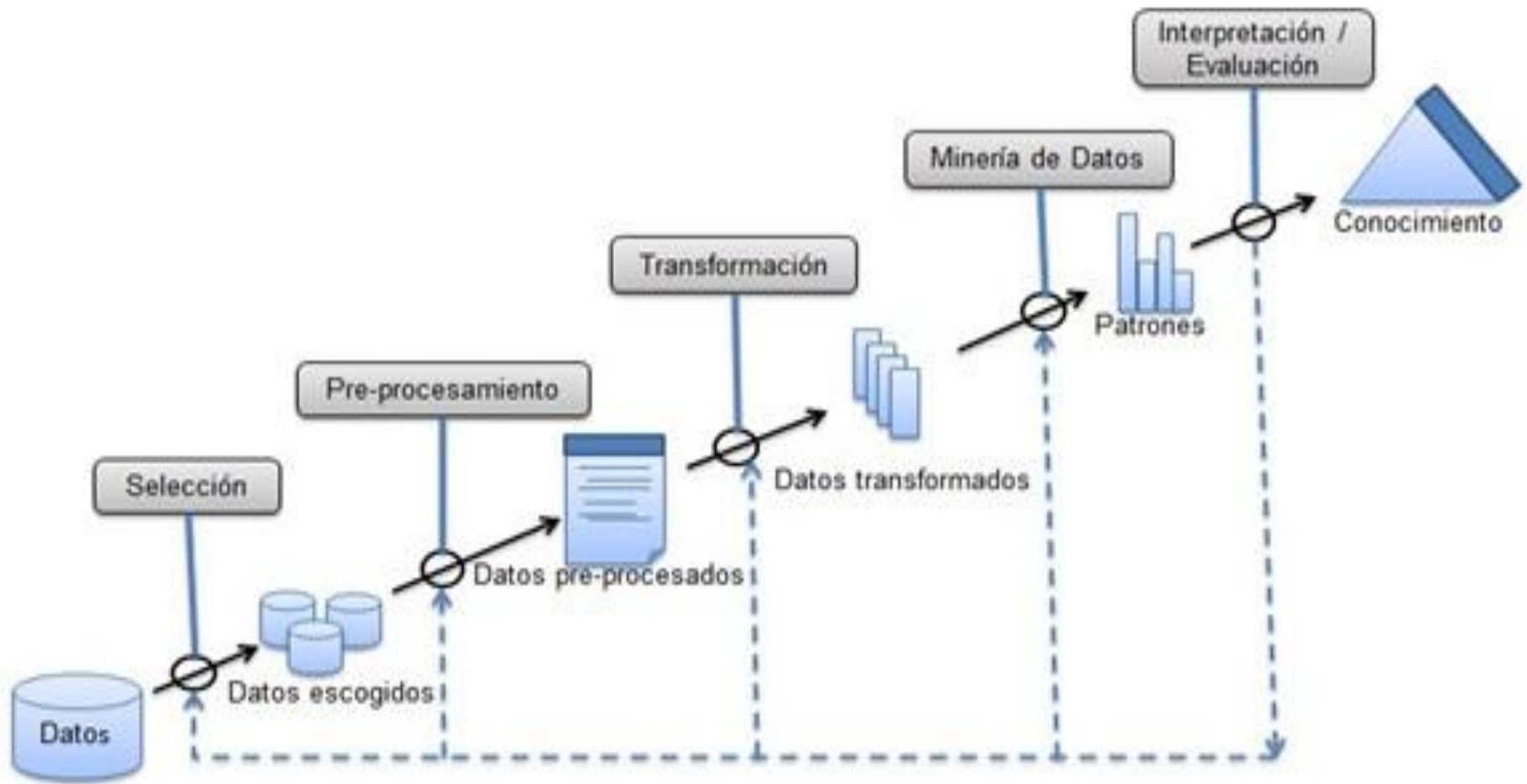
Trabajo de Diploma para optar por el título de Ingeniero en Informática

Autor: Ariam Rivas Méndez

Tutor: Ing. Adrian Pino Angulo

Curso: 2013-2014

Esquema general del proceso KDD



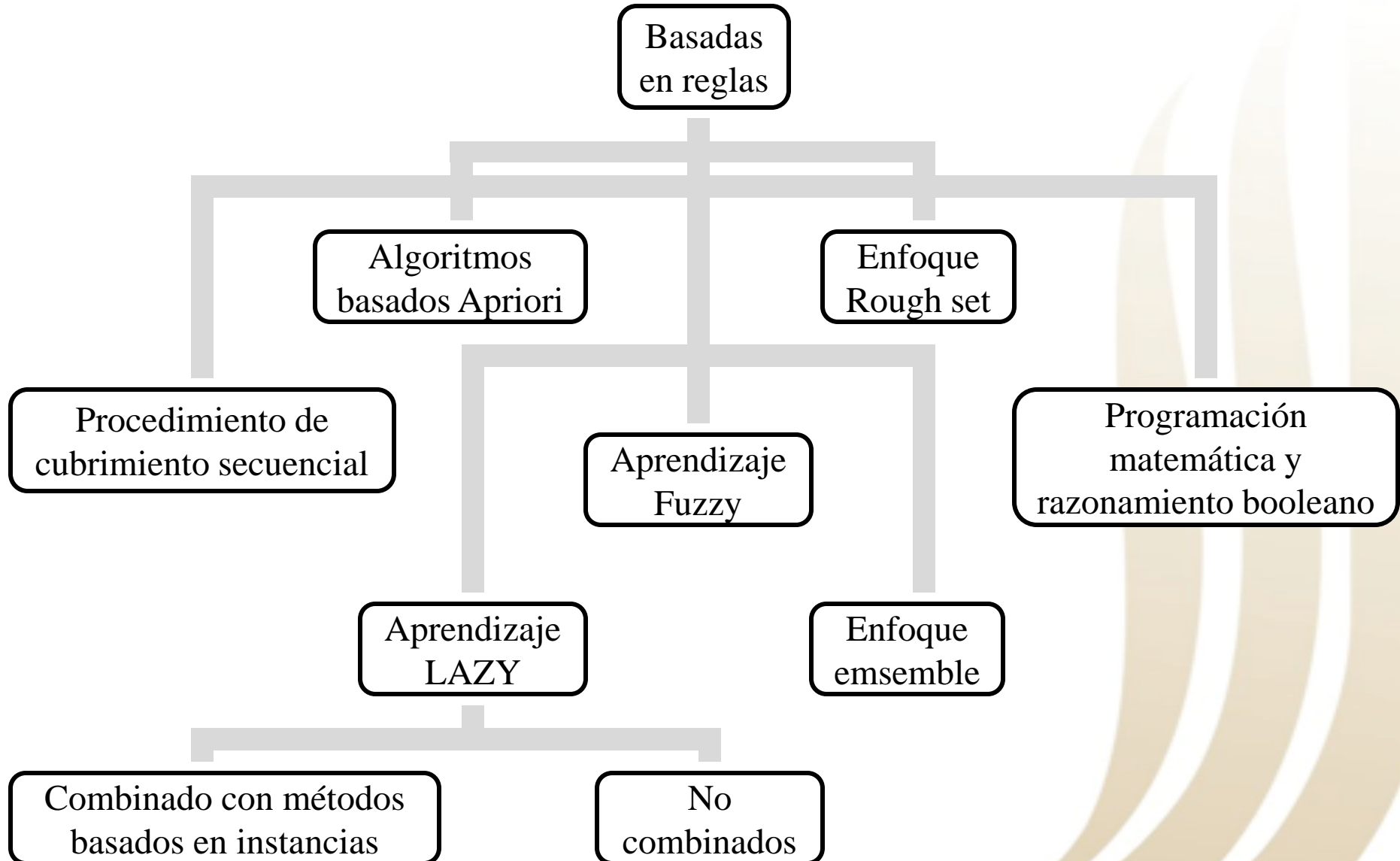
Conjuntos de datos de alta dimensión

- Conjuntos de datos definidos por miles de variables
- Requieren una gran cantidad de dimensiones desde el punto de vista algebraico para ser representados
- Afectan drásticamente la eficacia y eficiencia de los algoritmos de clasificación

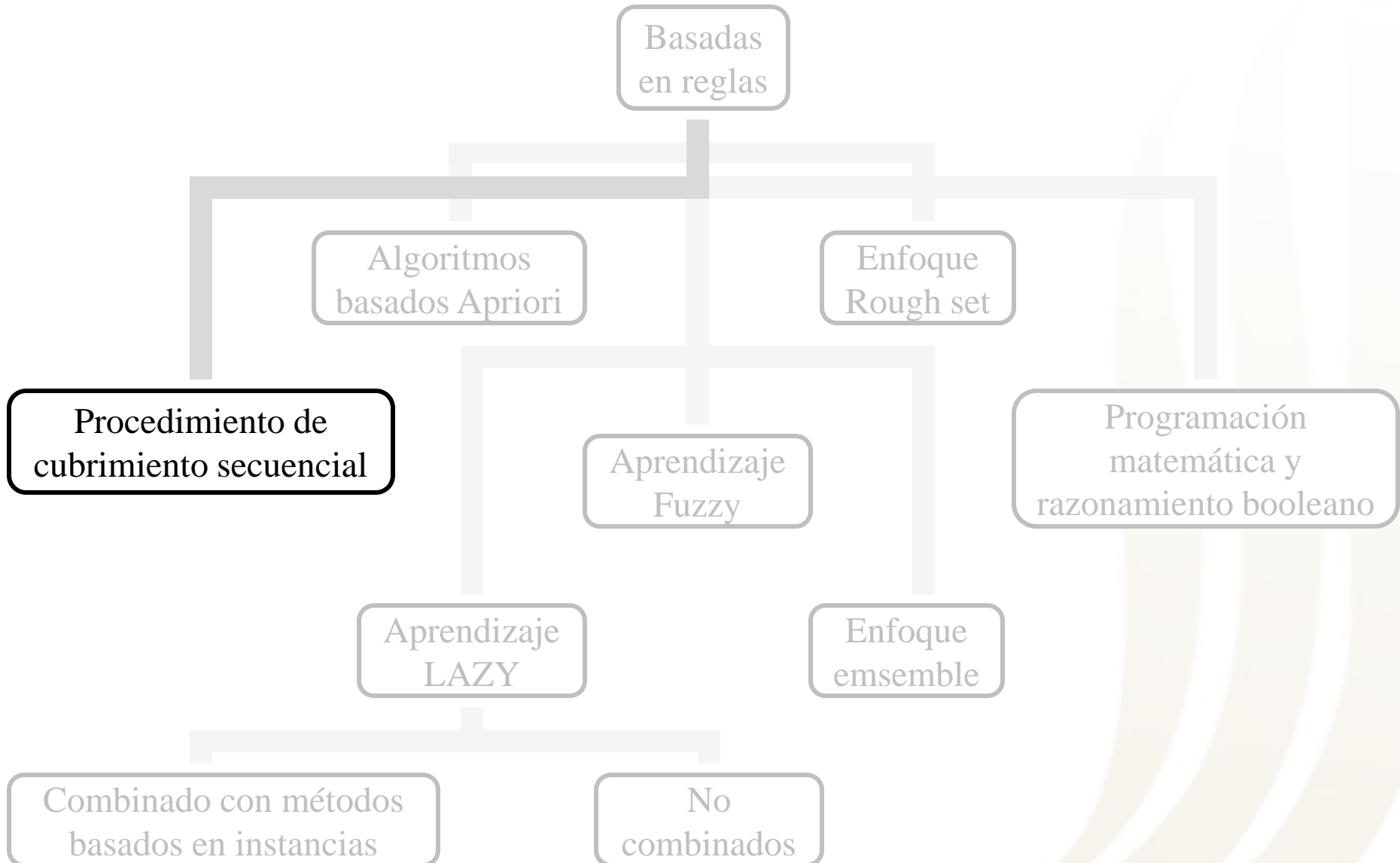
Ventaja de los algoritmos de clasificación basados en reglas

- Fácil comprensión por la estructura tan simple que poseen
- Pueden clasificar nuevas instancias rápidamente
- Pueden manejar fácilmente valores faltantes y atributos numéricos
- Su rendimiento es alto con respecto a otros algoritmos de clasificación

Taxonomía de algoritmos de clasificación basados en reglas



Taxonomía de algoritmos de clasificación basados en reglas



Deficiencias de dominio

Existe escasa evidencia, para decidir cuál de estos algoritmos es más propicio para conjuntos de datos de alta dimensión.

Esto generalmente conlleva a que se tenga que invertir mucho tiempo de experimentación y puede llevar a la selección inadecuada de un algoritmo de aprendizaje.

Problema

¿Cómo determinar la efectividad de los clasificadores basados en cubrimiento secuencial?

Objeto de estudio

El proceso de minería de datos a partir de clasificadores basados en cubrimiento secuencial

Objetivo

Realizar un estudio experimental de los algoritmos de clasificación basados en cubrimiento secuencial para la determinación de reglas, para su elección y aplicación

Campo de acción

Clasificadores basados en cubrimiento secuencial para la determinación de reglas

Preguntas científicas

¿Cuáles son las bases teóricas que sustentan los algoritmos de clasificación basados en reglas?

¿En qué radican los fenómenos que afectan la minería en datos de alta dimensión?

¿Qué características identifican a las principales técnicas de clasificación basadas en reglas?

Preguntas científicas

¿En qué grado están disponibles los algoritmos de aprendizaje basados en reglas más destacados?

¿Cómo realizar un estudio experimental que ofrezca la evidencia necesaria para enfrentar el problema?

¿Cuán efectivos resultan los principales métodos de clasificación basados en reglas ante conjuntos de datos de alta dimensión?

Tareas científicas

1. Estudiar las bases teóricas que sustentan el aprendizaje automático para la construcción de modelos de clasificación basados en reglas
2. Detallar los fenómenos que inciden en la minería de datos de alta dimensión y las estrategias para enfrentarlos
3. Confeccionar una taxonomía de los algoritmos de clasificación

Tareas científicas

4. Describir los algoritmos de aprendizaje basados en reglas para los cuales se reportan los resultados más destacados
5. Seleccionar una herramienta de minería de datos e incluir los algoritmos destacados que no se encuentren en la misma

Tareas científicas

6. Realizar un estudio experimental comparativo entre las técnicas más destacadas para hacer este tipo de minería de datos
7. Evaluar la efectividad de los principales métodos de clasificación basados en reglas ante conjuntos de datos de alta dimensión

Métodos de investigación científica

Teóricos

- Análisis y síntesis
- Histórico lógico
- Enfoque sistémico estructural

Empíricos

- Revisión de documentos
- Experimentación

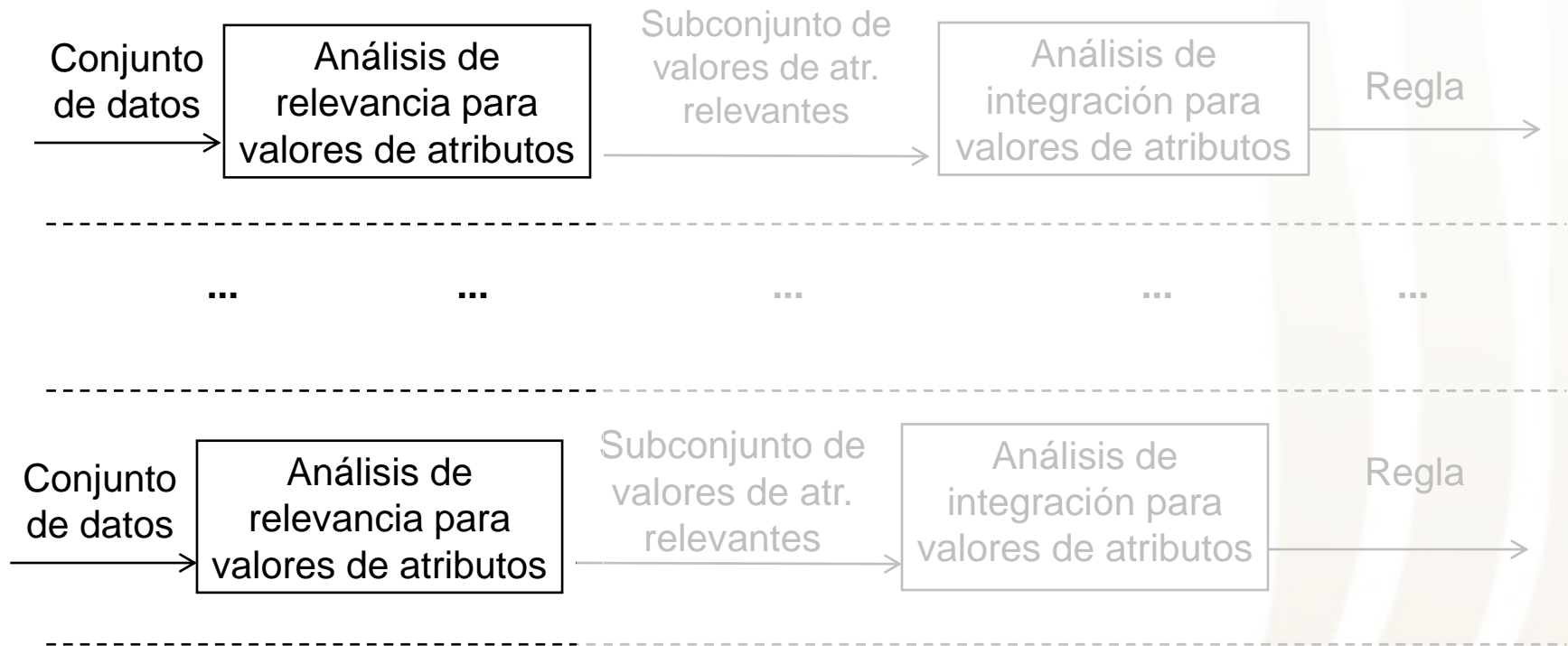
Estadísticos y Matemáticos

- Prueba no paramétricas de Friedman
- Prueba post-hoc de Holm

Implementación y mejora de nuevos algoritmos

PAVICD, ZigZag y ORIPPERk

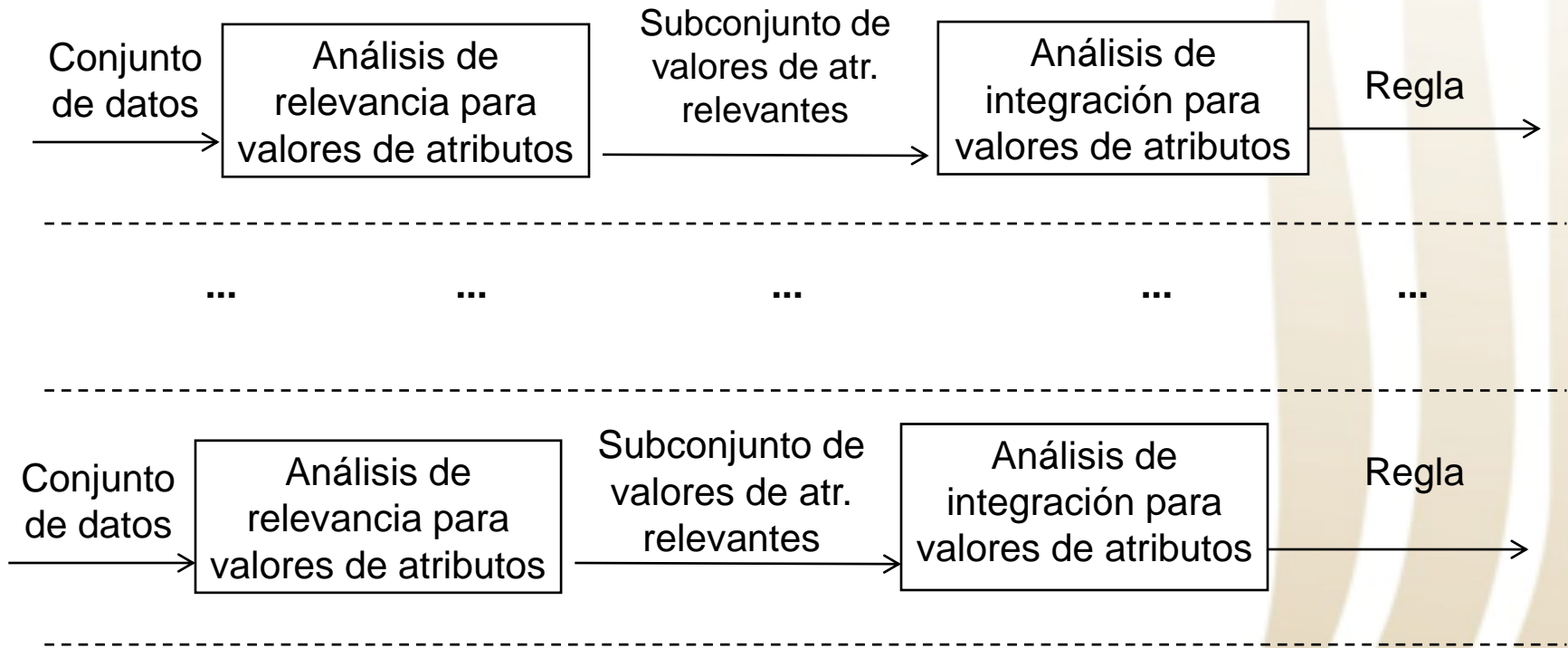
Para valor de clase C_1



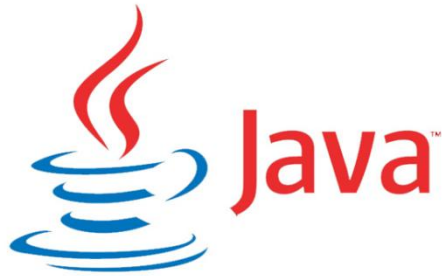
Implementación y mejora de nuevos algoritmos

PAVICD, ZigZag y ORIPPERk

Para valor de clase C_1



Software utilizados



Java



IDE NetBeans



Framework Weka 3.7.9

Clasificadores a evaluar

- One Rule (OneR)
- Ripple Down Rules (Ridor)
- Decision Table (DT)
- RIPPERk (Jrip)
- PART
- ORIPPERk
- PAVICD
- ZigZag

Clasificadores a evaluar

- One Rule (OneR)
- Ripple Down Rules (Ridor)
- Decision Table (DT)
- RIPPERk (Jrip)
- PART
- ORIPPERk
- PAVICD
- ZigZag

Parámetros a evaluar

- Precisión (accuracy)
- Tiempo de ejecución
- Número de condiciones de las reglas

Esquema de experimentación

- Validación cruzada con diez particiones y una corrida
- Pruebas no paramétricas (Friedman y Holm)
 $\alpha = 0.05$

Conjunto de datos

Conjunto de datos	Siglas	Atributos	Instancias	Clases
HEPATITIS	HEP	20	155	2
DERMATOLOGY	DER	35	366	6
ARRHYTHMIA	ARR	280	452	13
ADA	ADA	49	4147	2
OPTDIGITS	OPT	65	5620	10
SECOM	SEC	591	1567	2
SYLVA	SYL	217	13086	2
MULTIPLE FEATURES	MFE	650	2000	10
GINA	GIN	970	3153	2
ADS	ADS	1559	3279	2
HIVA	HIV	1618	3845	2
HEPATITISC	HPC	22278	123	4
ARCENE	ARC	10001	100	2
DOROTHEA	DOR	100001	800	2
DEXTER	DEX	20001	300	2
BURKITTLYMPHOMA	BLY	22284	220	3
ISOLET	ISO	618	6238	26

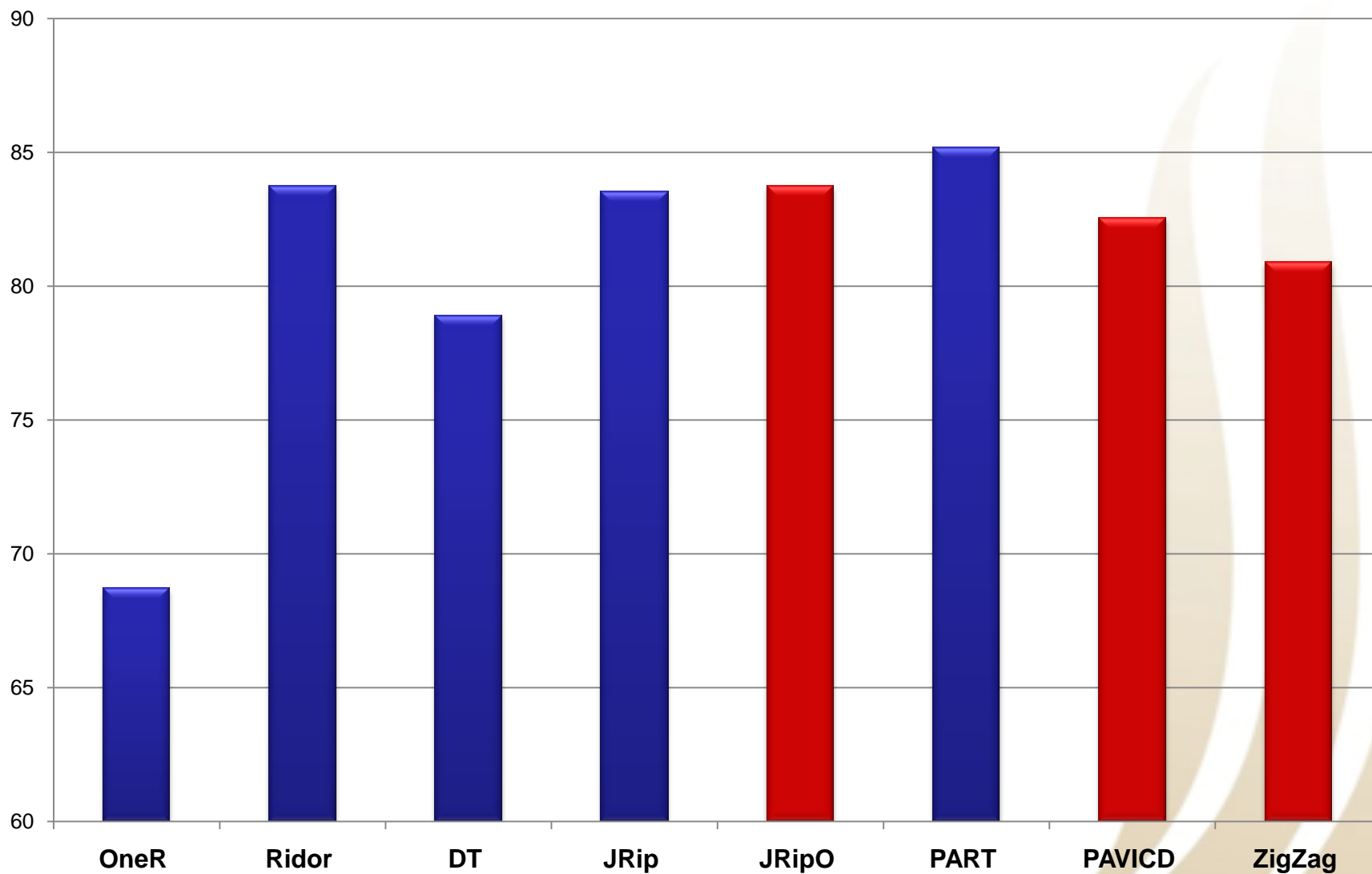
Discusión de los resultados



Precisión

DataSet	OneR	Ridor	DT	JRip	ORIPPERk	PART	PAVICD	ZigZag
HEP	83,23	78,71	76,13	78,06	78,06	84,52	81,94	81,29
DER	49,73	93,17	86,89	86,89	87,98	94,54	90,71	91,26
ARR	57,52	68,36	65,71	70,80	71,46	63,94	70,13	73,67
ADA	79,46	82,81	84,30	83,87	83,87	83,70	80,59	79,87
OPT	26,99	90,23	62,24	90,78	90,98	92,54	51,89	54,36
SEC	92,98	93,36	93,24	92,28	92,28	91,64	93,49	93,49
SYL	94,76	98,82	99,12	98,85	98,85	99,04	98,54	97,42
MFE	48,35	93,05	82,70	91,80	92,05	94,65	86,45	83,75
GIN	71,74	87,38	83,48	88,93	88,93	87,95	78,97	77,23
ADS	92,86	96,80	95,88	97,04	97,04	96,86	96,83	94,33
HIV	96,67	96,18	95,89	96,41	96,41	95,68	96,41	96,49
HPC	65,85	79,67	69,92	67,48	68,29	78,86	80,49	74,80
ARC	48,72	58,97	58,97	46,15	46,15	43,59	89,74	84,62
DOR	89,00	90,25	82,00	90,00	90,00	86,38	84,88	88,50
DEX	71,33	85,00	81,17	85,33	85,33	88,17	90,00	76,67
BLY	83,18	77,27	76,82	78,18	78,18	80,91	85,91	78,64
ISO	15,26	53,25	47,02	77,24	77,65	84,95	46,60	48,45

Precisión (Promedio)



Ranking Promedio - Prueba de Friedman

Algoritmo	Ranking
PART	3,65
ORIPPERk	3,76
PAVICD	4,03
Ridor	4,03
JRip	4,15
ZigZag	4,62
DT	5,47
OneR	6,29

P-valor: 0.0172 Precisión

Prueba de Holm para la Precisión

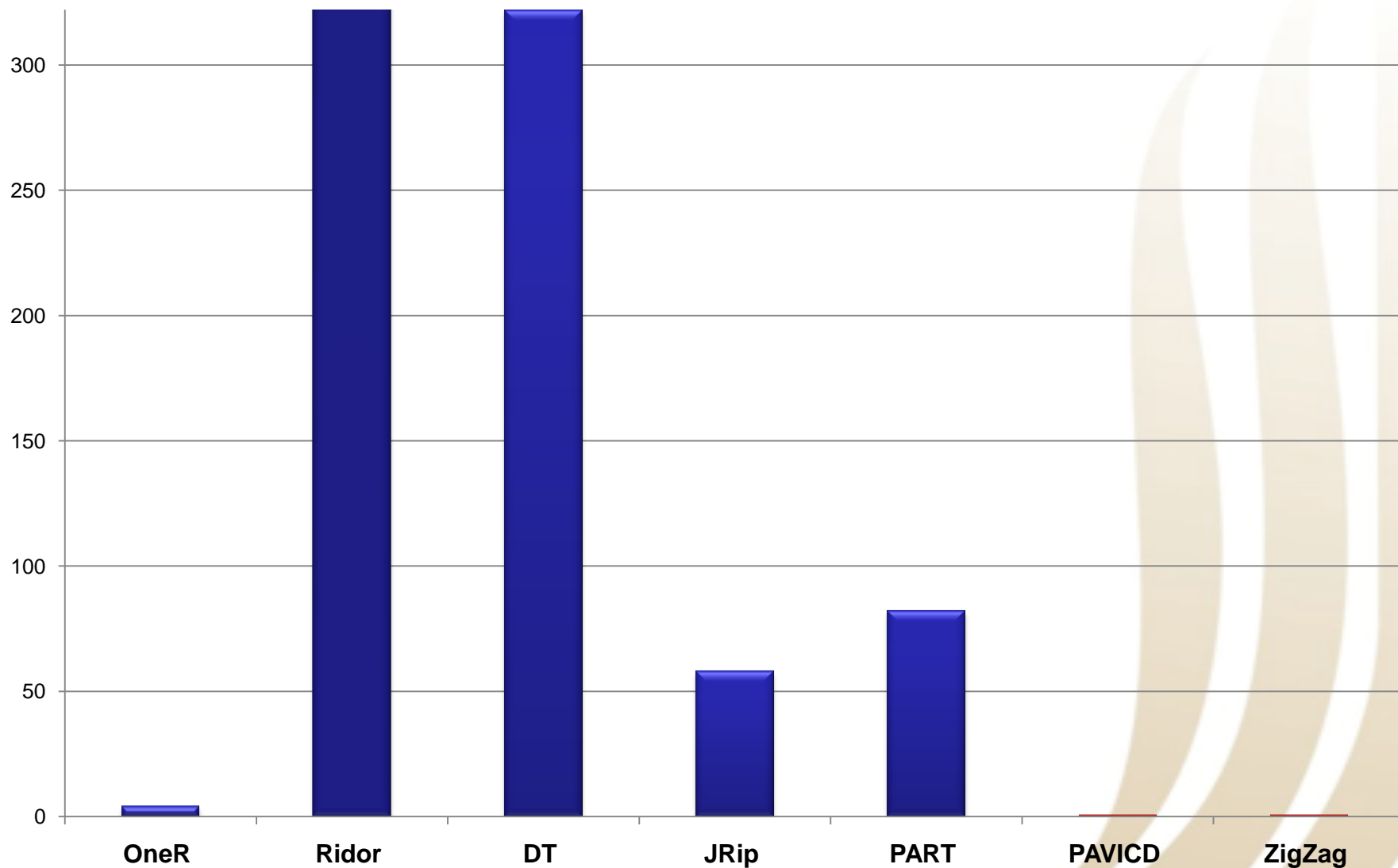
hypothesis	p_{Holm}
PART vs .OneR	4,56E-02
OneR vs .ORIPPERk	7,04E-02
OneR vs .PAVICD	1,83E-01
OneR vs .Ridor	1,83E-01
OneR vs .JRip	2,55E-01
DT vs .PART	6,89E-01
DT vs .ORIPPERk	9,31E-01
OneR vs .ZigZag	9,66E-01
DT vs .PAVICD	1,73E+00
Ridor vs .DT	1,73E+00
DT vs .JRip	2,07E+00
PART vs .ZigZag	4,22E+00
DT vs .ZigZag	4,96E+00
ZigZag vs .ORIPPERk	4,96E+00

hypothesis	p_{Holm}
OneR vs .DT	4,96E+00
PAVICD vs .ZigZag	6,29E+00
Ridor vs .ZigZag	6,29E+00
JRip vs .PART	6,29E+00
JRip vs .ZigZag	6,29E+00
Ridor vs .PART	6,29E+00
JRip vs .ORIPPERk	6,29E+00
PART vs .PAVICD	6,29E+00
Ridor vs .ORIPPERk	6,29E+00
PAVICD vs .ORIPPERk	6,29E+00
JRip vs .PAVICD	6,29E+00
Ridor vs .JRip	6,29E+00
PART vs .ORIPPERk	6,29E+00
Ridor vs .PAVICD	6,29E+00

Tiempo de ejecución

DataSet	OneR	Ridor	DT	JRip	PART	PAVICD	ZigZag
HEP	0,01	0,01	0,10	0,06	0,06	0,03	0,00
DER	0,00	0,04	0,08	0,02	0,01	0,01	0,00
ARR	0,01	2,92	1,32	0,44	0,84	0,03	0,01
ADA	0,01	0,29	0,98	0,35	1,00	0,01	0,01
OPT	0,05	61,87	4,25	4,37	1,51	0,01	0,02
SEC	0,14	0,81	2,83	1,60	3,82	0,05	0,05
SYL	0,25	2,22	76,61	8,14	3,79	0,08	0,09
MFE	0,17	27,89	8,30	7,34	3,72	0,09	0,08
GIN	0,54	9,93	41,43	18,71	22,39	0,19	0,19
ADS	0,35	5,63	133,29	9,03	48,83	0,32	0,32
HIV	0,40	3,47	66,03	6,77	17,34	0,38	0,38
HPC	0,21	4,31	25,88	7,59	3,24	0,25	0,22
ARC	0,02	0,11	3,48	0,19	0,15	0,04	0,04
DOR	62,45	375,76	4849,32	618,37	1026,87	7,41	7,68
DEX	1,30	12,20	141,40	29,92	23,35	0,79	0,80
BLY	0,53	5,23	39,03	13,73	7,93	0,34	0,32
ISO	0,94	345600,00	72,67	254,46	230,25	0,24	0,24

Tiempo de ejecución (Promedio)



Ranking Promedio - Prueba de Friedman

Algoritmo	Ranking
ZigZag	1,65
PAVICD	1,97
OneR	2,5
Ridor	4,79
PART	5,24
JRip	5,38
DT	6,47

P-valor: 5.107E-11 Tiempo

Prueba de Holm para Tiempo de ejecución

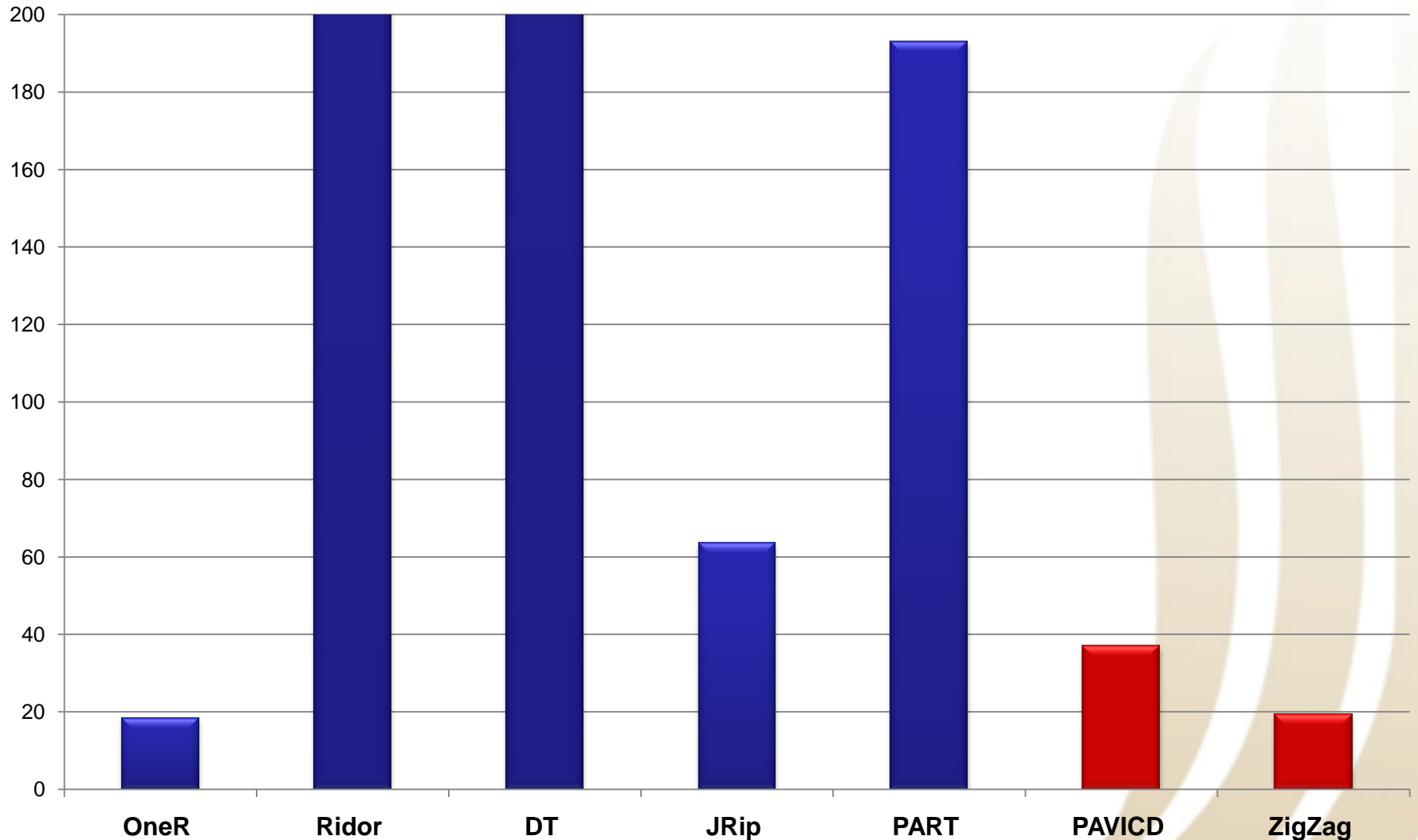
Hypothesis	Holm (Adjusted p-values)
ZigZag vs .DT	1,57974E-09
PAVICD vs .DT	2,50750E-08
OneR vs .DT	1,59247E-06
ZigZag vs .JRip	8,32873E-06
ZigZag vs .PART	2,17754E-05
PAVICD vs .JRip	6,61443E-05
PAVICD vs .PART	1,57904E-04
ZigZag vs .Ridor	3,02934E-04
OneR vs .JRip	1,30300E-03
PAVICD vs .Ridor	1,66326E-03
OneR vs .PART	2,45170E-03

Hypothesis	Holm (Adjusted p-values)
OneR vs .Ridor	1,96052E-02
Ridor vs .DT	2,12960E-01
DT vs .PART	7,63866E-01
DT vs .JRip	9,93426E-01
ZigZag vs .OneR	1,49807E+00
Ridor vs .JRip	2,13631E+00
PAVICD vs .OneR	2,13631E+00
Ridor vs .PART	2,13631E+00
ZigZag vs .PAVICD	2,13631E+00
JRip vs .PART	2,13631E+00

Número de condiciones de las reglas

DataSet	OneR	Ridor	DT	JRip	PART	PAVICD	ZigZag
HEP	3	1	135	7	20	5	4
DER	4	11	380	20	19	29	25
ARR	5	534	396	22	156	76	54
ADA	6	144	984	26	1038	10	4
OPT	11	19750	2500	305	423	54	36
SEC	4	1	10	8	47	6	4
SYL	25	20	5317	27	226	40	6
MFE	62	268	804	64	61	43	33
GIN	12	37	4407	77	227	31	9
ADS	23	18	1653	25	97	43	6
HIV	2	5	1800	20	231	17	8
HPC	4	8	84	9	8	11	9
ARC	3	3	18	2	4	5	10
DOR	2	1	1587	1	44	98	6
DEX	4	11	414	25	35	58	8
BLY	3	4	96	7	10	24	18
ISO	139	72	4782	434	633	79	86

Número de condiciones de las reglas (Promedio)



Ranking Promedio - Prueba de Friedman

Algoritmo	Ranking
OneR	2,06
ZigZag	2,82
Ridor	3,03
JRip	3,88
PAVICD	4,35
PART	5,09
DT	6,76

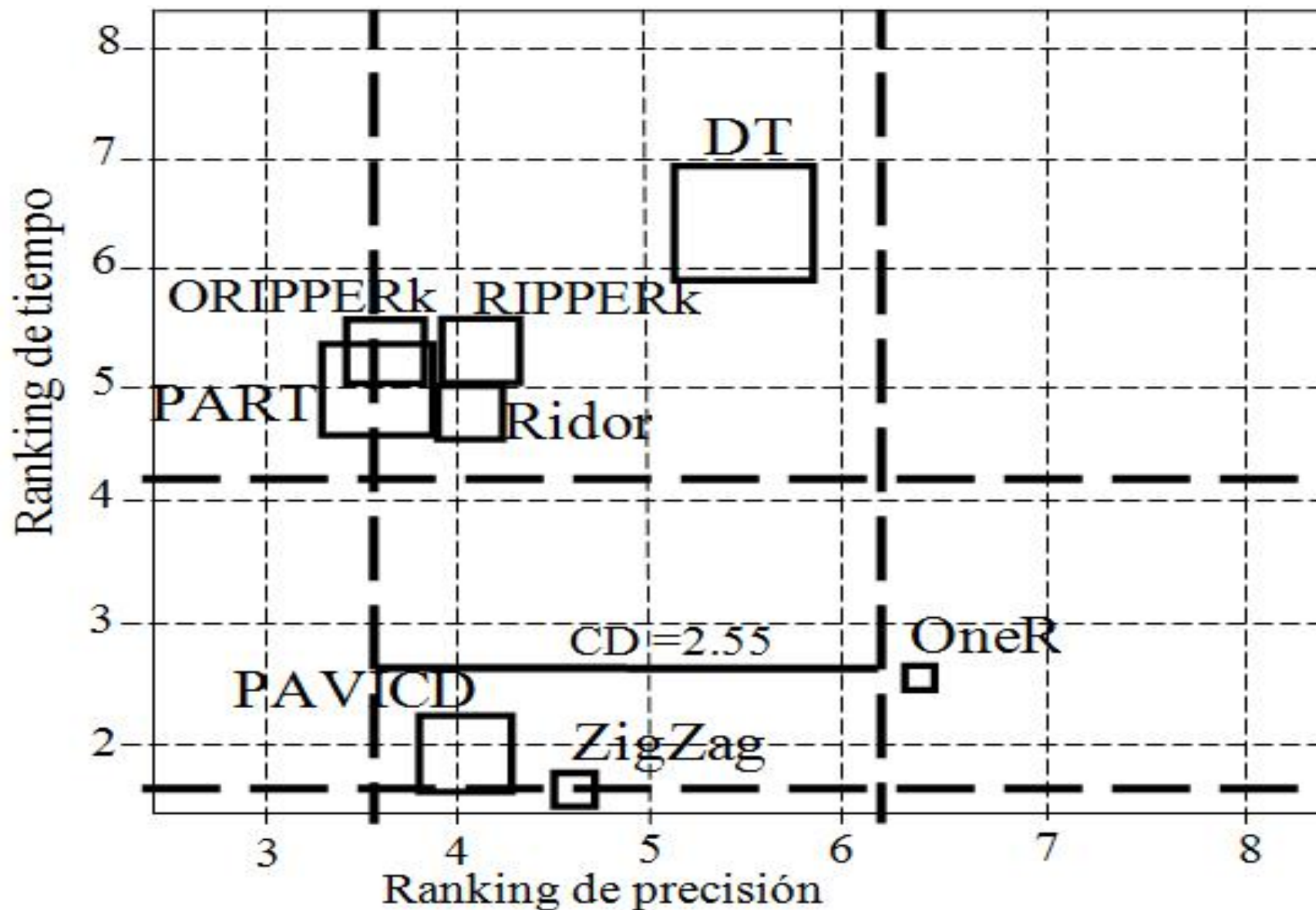
P-valor: 5.453E-10 Número de Condiciones

Prueba de Holm para el número de condiciones

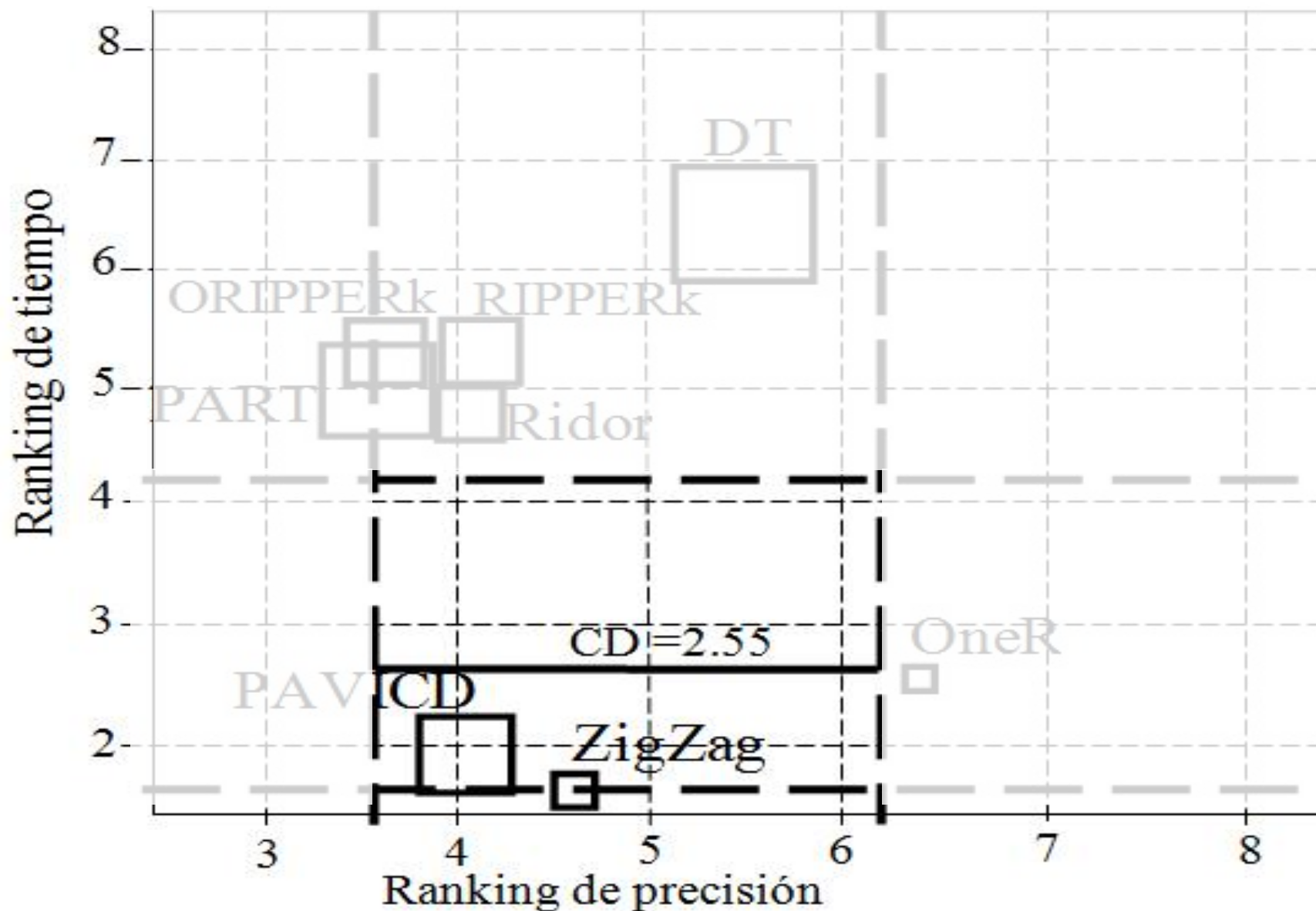
Hypothesis	Holm (Adjusted p-values)
OneR vs .DT	4,490E-09
ZigZag vs .DT	2,087E-06
Ridor vs .DT	8,791E-06
OneR vs .PART	7,815E-04
JRip vs.DT	1,704E-03
PAVICD vs .DT	1,815E-02
OneR vs .PAVICD	2,941E-02
ZigZag vs .PART	3,136E-02
Ridor vs .PART	7,097E-02
OneR vs .JRip	1,662E-01
DT vs .PART	2,603E-01

Hypothesis	Holm (Adjusted p-values)
ZigZag vs PAVICD	3,901E-01
Ridor vs .PAVICD	6,665E-01
JRip vs .PART	8,291E-01
ZigZag vs .JRip	1,071E+00
OneR vs .Ridor	1,141E+00
Ridor vs .JRip	1,248E+00
OneR vs .ZigZag	1,248E+00
PAVICD vs .PART	1,248E+00
JRip vs .PAVICD	1,248E+00
ZigZag vs .Ridor	1,248E+00

Ranking de precisión, tiempo y número de condiciones (Test Nemenyi)



Ranking de precisión, tiempo y número de condiciones (Test Nemenyi)



Conclusiones

- Se describieron los algoritmos más reconocidos en el área de estudio de los clasificadores basados en cubrimiento secuencial
- Se implementó el algoritmo PAVICD y ZigZag y se mejoró el algoritmo RIPPERk
- Se realizó una experimentación de los algoritmos de clasificación detallados en este estudio

Conclusiones

- Los algoritmos más rápidos resultaron ser PAVICD y ZigZag y mostraron su competitividad en cuanto a la precisión alcanzada
- Se constató que el algoritmo PART supera significativamente a OneR en cuanto a precisión
- El algoritmo Decision Table genera significativamente mayor número de condiciones con respecto al resto de los algoritmos (exceptuando PART)

Recomendaciones

- Estudiar y describir otros algoritmos basados en cubrimiento secuencial
- Extender este estudio experimental utilizando otros conjuntos de datos
- Mejorar los algoritmos PAVICD y ZigZag de manera que obtengan reglas con mayor precisión

ESTUDIO EXPERIMENTAL SOBRE ALGORITMOS DE CLASIFICACIÓN SUPERVISADA BASADOS EN CUBRIMIENTO SECUENCIAL

Trabajo de Diploma para optar por el título de Ingeniero en Informática

Autor: Ariam Rivas Méndez

Tutor: Ing. Adrian Pino Angulo

Curso: 2013-2014

Pregunta 1

¿A qué responde la selección de las pruebas estadísticas no paramétricas?

Respuesta

Requerimientos de las pruebas paramétricas:

- Los datos deben seguir una distribución normal
- Los datos deben presentar homogeneidad respecto a su varianza

En el entorno del aprendizaje automático se violan tales requerimientos al comparar el desempeño de los algoritmos sobre los conjuntos de datos.

Respuesta

Se recomienda el uso de pruebas no paramétricas para comparar algoritmos según:

- precisión
- tasas de error
- tamaños del modelo
- tiempo de ejecución

Pregunta 2

En un número importante de los conjuntos de datos empleados para la comparación de los algoritmos se realiza una clasificación binaria, a la vez que varios de ellos presentan desbalance respecto a la cantidad de clases. ¿Cómo influye esta situación en la competitividad de los algoritmos propuestos (PAVICD y ZigZag) atendiendo a la precisión? Argumente.

Precisión

DataSet	OneR	Ridor	DT	JRip	ORIPPERk	PART	PAVICD	ZigZag
HEP	83,23	78,71	76,13	78,06	78,06	84,52	81,94	81,29
DER	49,73	93,17	86,89	86,89	87,98	94,54	90,71	91,26
ARR	57,52	68,36	65,71	70,80	71,46	63,94	70,13	73,67
ADA	79,46	82,81	84,30	83,87	83,87	83,70	80,59	79,87
OPT	26,99	90,23	62,24	90,78	90,98	92,54	51,89	54,36
SEC	92,98	93,36	93,24	92,28	92,28	91,64	93,49	93,49
SYL	94,76	98,82	99,12	98,85	98,85	99,04	98,54	97,42
MFE	48,35	93,05	82,70	91,80	92,05	94,65	86,45	83,75
GIN	71,74	87,38	83,48	88,93	88,93	87,95	78,97	77,23
ADS	92,86	96,80	95,88	97,04	97,04	96,86	96,83	94,33
HIV	96,67	96,18	95,89	96,41	96,41	95,68	96,41	96,49
HPC	65,85	79,67	69,92	67,48	68,29	78,86	80,49	74,80
ARC	48,72	58,97	58,97	46,15	46,15	43,59	89,74	84,62
DOR	89,00	90,25	82,00	90,00	90,00	86,38	84,88	88,50
DEX	71,33	85,00	81,17	85,33	85,33	88,17	90,00	76,67
BLY	83,18	77,27	76,82	78,18	78,18	80,91	85,91	78,64
ISO	15,26	53,25	47,02	77,24	77,65	84,95	46,60	48,45

Precisión

DataSet	Ridor	ORIPPERk	PART	PAVICD	ZigZag
HEP	78,71	78,06	84,52	81,94	81,29
ADA	82,81	83,87	83,70	80,59	79,87
SEC	93,36	92,28	91,64	93,49	93,49
SYL	98,82	98,85	99,04	98,54	97,42
ADS	96,80	97,04	96,86	96,83	94,33
HIV	96,18	96,41	95,68	96,41	96,49
DOR	90,25	90,00	86,38	84,88	88,50

Respuesta

Algoritmo	Ranking
ORIPPERk	2,50
PART	2,86
Ridor	3,14
PAVICD	3,14
ZigZag	3,36

P-valor	
Friedman	0,87
Iman-Daveport	0,89

Respuesta

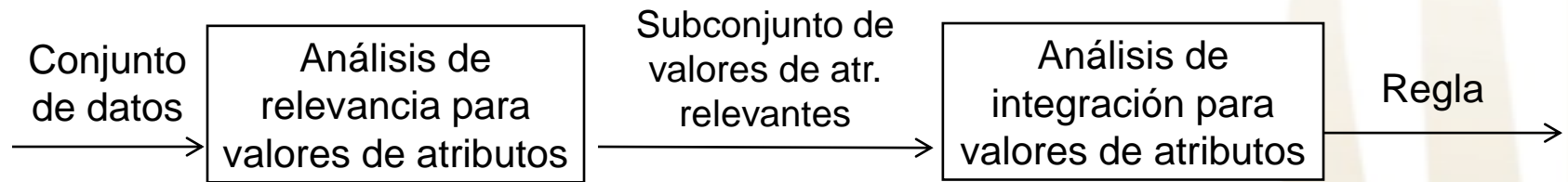
ORIPPERk, PART y Ridor crean una única regla enfocada a la clase menos presente en el conjunto de datos, de modo que a todas aquellas instancias que no cumplan con dicha regla, se les asigna la clase alternativa.

Pregunta 3

¿Considera usted que las bases teóricas de los algoritmos propuestos (PAVICD y ZigZag) pudieran emplearse para mejorar otros clasificadores basados en reglas y que también resulten procedimientos de cubrimiento secuencial? Explique.

Respuesta

Para valor de clase C_1



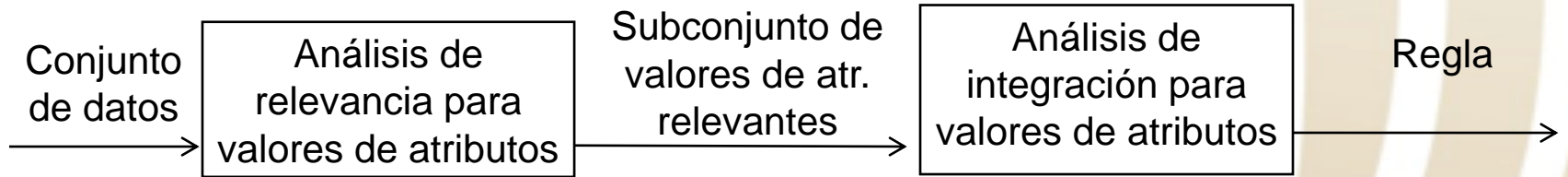
...

...

...

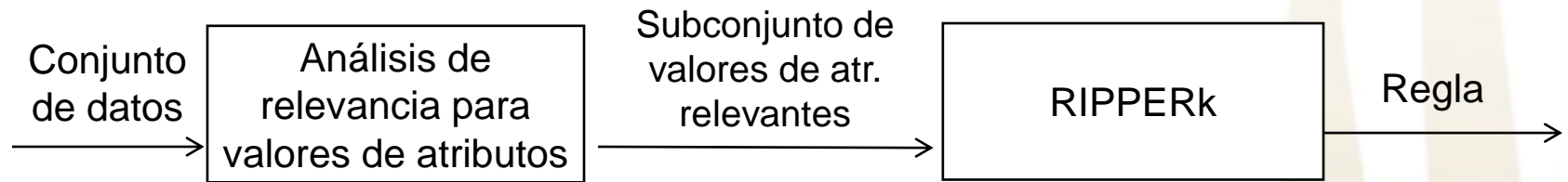
...

...



Respuesta

Para valor de clase C_1



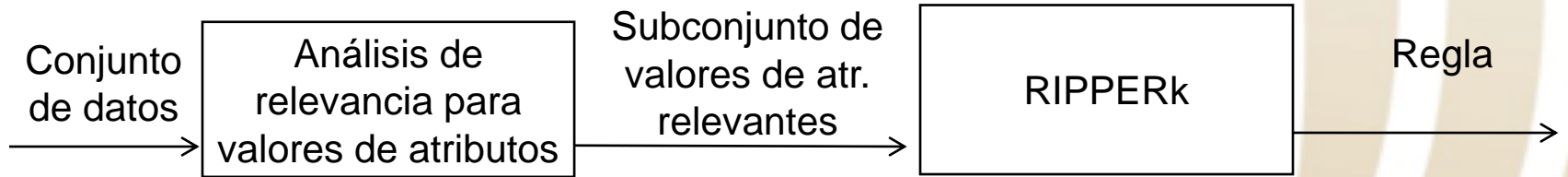
...

...

...

...

...



Pregunta 4

Argumente con qué propósitos fueron planteadas las siguientes líneas de trabajo futuro:

- "Extender este estudio experimental utilizando otros conjuntos de datos."
- "Mejorar los algoritmos PAVICD y ZigZag de manera que obtengan reglas con mayor precisión."

Pregunta 4

Argumente con qué propósitos fueron planteadas las siguientes líneas de trabajo futuro:

- "Extender este estudio experimental utilizando otros conjuntos de datos."
- "Mejorar los algoritmos PAVICD y ZigZag de manera que obtengan reglas con mayor precisión."

Respuesta

- Se pretende extraer conocimiento de conjuntos de datos de dominios específicos:

ej. el entorno médico incluye aplicaciones bioquímicas, estudio de fármacos... donde los conjuntos de datos suelen ser, por naturaleza, de alta dimensión.

- Aumentar el número de conjuntos de datos debe mejorar la confianza en los resultados del análisis estadístico.

Pregunta 4

Argumente con qué propósitos fueron planteadas las siguientes líneas de trabajo futuro:

- "Extender este estudio experimental utilizando otros conjuntos de datos."
- "Mejorar los algoritmos PAVICD y ZigZag de manera que obtengan reglas con mayor precisión."

Respuesta

$$R(A_i^j; C_k) = \underbrace{\alpha_k P(A_i^j | C_k)}_{\text{Cubrimiento}} + \underbrace{(1 - \alpha_k) P(C_k | A_i^j)}_{\text{Confiabilidad}}$$

Mientras (condición de parada){

1. Paso

Determinar la clase crítica:

$$C_q = \max_{\forall C_k \in C} \{ |InCR_{C_k} - InCR_{C_{Tk}}| \}$$

2. Paso

Actualizar el valor de α para la clase C_q

3. Paso

Integrar los valores de atributos de la Clase C_q .
}

	C_0	C_1	C_2	C_3
C_0	10	2	1	4
C_1	0	8	3	2
C_2	0	1	12	1
C_3	2	1	5	6

→ Instancias no cubiertas por la regla

→ Instancias de otra clase cubiertas por la regla

	C_0	C_1	C_2	C_3
C_0	13	1	0	3
C_1	0	8	3	2
C_2	1	0	12	1
C_3	2	1	5	6

ESTUDIO EXPERIMENTAL SOBRE ALGORITMOS DE CLASIFICACIÓN SUPERVISADA BASADOS EN CUBRIMIENTO SECUENCIAL

Trabajo de Diploma para optar por el título de Ingeniero en Informática

Autor: Ariam Rivas Méndez

Tutor: Ing. Adrian Pino Angulo

Curso: 2013-2014