Apéndice A

Bases de datos con variable de clasificación crisp

Para la confección de la parte experimental de esta memoria, se han seleccionado algunas bases de datos. Estas bases de datos reflejan los diversos problemas a los que se puede enfrentar un algoritmo de aprendizaje.

De cada base de datos discribiremos cada uno de los atributos. Para cada atributo aportaremos la siguiente información:

- **Variable**: Identificador utilizada para identificar a dicho atributo en la descripción de las reglas.
- **Referencia**: Tipo del universo de discurso de la variable que puede tomar únicamente tres valores: Nominal, Discreto o Continuo.
- **Rango**: Describe el universo de discurso de las variables con referencial discreto o continuo. En el caso de que el atributo sea nominal, se describe el número de valores que dicho atributo puede tomar.
- **Val. perdidos**: Indica si este atributo contiene valores perdidos.
- **N. etiquetas**: Indica el número de etiquetas difusas definidas para discretizar los universos de discurso de las variables continuas o discretas. En el caso de variables nominales, este valor coincide con mismo asignado al rango. Para el caso discreto y continuo, asumiremos que el número de etiquetas se distribuye de forma uniforme sobre el rango de definición cortandose a grado 0.5.

La bases de datos que más adelante pasaremos a describir, se han obtenido del UCI Respository of Machine Learning Databases and Domain Theories, cuya dirección ftp es *ftp.ics.uci.edu*, excepto la base de datos INFARTO.

A continuación describimos las diferentes bases de datos utilizadas.

- **Base de datos IRIS**

  Descripción: Esta es una de las bases de datos más conocidas y utilizadas para mostrar el comportamiento de los algoritmos de clasificación. Fué propuesta en [38] y propone la clasificación de tres clases de plantas usando cuatro variables predictivas, usando 150 ejemplos.

  Para la parte experimental, se han construido 5 particiones en ficheros de entrenamiento y prueba. Dichas particiones se han realizado de forma completamenta aleatoria, de manera, que aproximadamente el 70 % de los ejemplos están en el conjunto de entrenamiento, y el 30 % en el conjunto de prueba.

  La variable de clasificación es nominal y puede tomar los tres siguientes valores: *Setosa, Virginica, Versicolor.*

Variable	Referencial	Rango	Val. perdidos	N. etiquetas
Sepal_l	Continuo	[4,8]	No	5
Sepal_w	Continuo	[2,4.5]	No	5
Petal_l	Continuo	[1,7]	No	5
Petal_w	Continuo	[0,3]	No	5

Tabla A.1: Base de datos IRIS

- **Base de datos INFARTO**

  Descripción: Esta base de datos fué creada por la profesora Dña Aurora Valenzuela, miembro del Departamento de Medicina Legal y Forense de la Universidad de Granada. A su condición de profesora, une la de ser medico-forense del Hospital Clínico de la ciudad de Granada. Los datos representan análisis reales realizados por la propia doctora sobre pacientes. La base de datos contiene 14 variables para determinar una variable nominal de diagnóstico. Dos de las variables utilizadas son nominales, el resto tienen un referencial continuo. Las variables continuas se pueden dividir en dos diferentes de pruebas: métodos morfológicos y análisis bioquímicos. Estas pruebas se aceptan como los mejores marcadores para el diagnóstico post-morten del infarto de miocardio. En algunos trabajos, tales como [52, 78], se ha utilizado esta base de datos. La base de datos INFARTO tiene 78 ejemplos y presenta valores perdidos.

  Para la parte experimental, se han construido 5 particiones en ficheros de entrenamiento y prueba. Dichas particiones se han realizado de forma completamenta aleatoria, de manera, que aproximadamente el 70 % de los ejemplos están en el conjunto de entrenamiento, y el 30 % en el conjunto de prueba.

  La variable de clasificación puede tomar dos valores: *INFARTO y NO-INFARTO.*

- **Bases de datos MONK**

Variable	Referencial	Rango	Val. perdidos	N. etiquetas
ApoA	Continuo	[0,140]	No	5
ApoB	Continuo	[0,140]	No	5
Lp(a)	Continuo	[0,2]	No	5
HDL	Continuo	[0,100]	No	5
LDL	Continuo	[0,150]	No	5
PARAOXONASA	Continuo	[0,200]	Si	5
SEXO	Nominal	2	No	2
EDAD	Discreto	[0,120]	No	5
MIOGLEB	Discreto	4	No	4
NTZ	Discreto	2	No	2
K/Na(A)	Continuo	[0,2]	No	5
K/Na(B)	Continuo	[0,2]	No	5
K/Na(C)	Continuo	[0,2]	No	5
K/Na(D)	Continuo	[0,2]	No	5

Tabla A.2: Base de datos INFARTO

Descripción: Los problemas del MONK fueron la base de la primera comparación internacional de algoritmos de aprendizaje. Los resultados de esta comparación se presentaron en [148]. Hay tres problemas del MONK. Los dominios para cada uno de ellos son los mismos, y además, se construyó una partición en ejemplos de prueba y en ejemplos de entrenamiento para cada uno de ellos.

Cada problema se compone de 432 ejemplos, seis variables discretas predicitivas, y una variable de clasificación, también discreta que toma dos valores: *0, 1*.

Variable	Referencial	Rango	Val. perdidos	N. etiquetas
a1	Discreto	3	No	3
a2	Discreto	3	No	3
a3	Discreto	2	No	2
a4	Discreto	3	No	3
a5	Discreto	4	No	4
a6	Discreto	2	No	2

Tabla A.3: Base de datos MONK

Cada problema representa un comportamiento diferente sobre los mismos atributos. El comportamiento que describen los ejemplos en cada uno de ellos es el siguiente:

– **MONK1**: El valor de la variable consecuente es 1, si a1=a2 o a5=1. En otro caso la variable consecuente toma valor 0.

– **MONK2**: La variable consecuente toma valor 1 si se satisfacen estrictamente dos de las siguientes condiciones: a1=1, a2=1, a3=1, a4=1, a5=1, a6=1. En otro caso el valor de la variable consecuente es 0.

– **MONK3**: El valor de la variable consecuente es 1, si (a5=3 y a4=1) o (a5≠4 y a2 ≠ 3). En otro caso la variable consecuente toma el valor 0. En esta base de datos, hay que destacar que se incluye un 5 % de ruido sobre el conjunto de entrenamiento.

- **Base de datos WINE**

  Descripción: Esta base de datos es el resultado de un análisis químico de vinos de un misma región con diferentes tipos de uva, usando 13 variables continuos y 178 ejemplos. Hemos usado 7 etiquetas lingüísticas uniformemente distribuidas para discretizar el rango de cada variable.

  La variable de clasificación es nominal, y puede tomar los siguientes tres valores: *Tipo1, Tipo2, Tipo3.*

  Para la parte experimental, se han construido 5 particiones en ficheros de entrenamiento y prueba. Dichas particiones se han realizado de forma completamenta aleatoria, de manera, que aproximadamente el 70 % de los ejemplos están en el conjunto de entrenamiento, y el 30 % en el conjunto de prueba.

Variable	Referencial	Rango	Val. perdidos	N. etiquetas
Att1	Continuo	[11,15]	No	7
Att2	Continuo	[0.74,5.8]	No	7
Att3	Continuo	[1.36,,3.23]	No	7
Att4	Continuo	[10.6,30]	No	7
Att5	Continuo	[70,162]	No	7
Att6	Continuo	[0.9,3.88]	No	7
Att7	Continuo	[0.34,5.08]	No	7
Att8	Continuo	[0.13,0.66]	No	7
Att9	Continuo	[0.41,3.58]	No	7
Att10	Continuo	[1.2,13]	No	7
Att11	Continuo	[0.48,1.71]	No	7
Att12	Continuo	[1.27,4]	No	7
Att13	Continuo	[278,1680]	No	7

Tabla A.4: Base de datos WINE

- **Base de datos PIMA**

  Descripción: Esta base de datos es una muestra extraída sobre una población cercana a Fenix, Arizona, EEUU sobre la diagnosis de la diabetes tomando como marco de referencia los signos propuestos por la Organización Mundial de la Salud. La base de datos consta de 768 ejemplos, dividos en 576 de entrenamiento y 192 de prueba. Esta base de datos se compone de 8 variables predictivas continuas más

Variable	Referencial	Rango	Val. perdidos	N. etiquetas
Pregnant	Continuo	[0,17]	No	5
Glucosa	Continuo	[0,199]	No	5
Presión	Continuo	[0,122]	No	5
Triceps	Continuo	[0,99]	No	5
Insulina	Continuo	[0,846]	No	5
Body_mass	Continuo	[0,67.1]	No	5
Pedigree	Continuo	[0.08,2.42]	No	5
Age	Discreto	[21,81]	No	5

Tabla A.5: Base de datos PIMA

una variable de clasificación binaria, donde el valor 1 se intrepreta como que el test de diabetes es positivo.

- **Base de datos IONOSPHERE**

  Descripción: Estos datos de radar fuerón recogidos por un sistema en Goose Bay, Labrador [141]. Este sistema consiste en un vector que contempla valores de 16 antenas de alta frecuencia con una potencia total transmitida del orden de 2.4 Kilowatios. El objetivo consistía en detectar electrones libres en la ionosfera. La base de datos está compuesta de 34 atributos continuos además de la variable de clasificación con 351 ejemplos.

  Para la parte experimental, se han construido 5 particiones en ficheros de entrenamiento y prueba. Dichas particiones se han realizado de forma completamenta aleatoria, de manera, que aproximadamente el 70 % de los ejemplos están en el conjunto de entrenamiento, y el 30 % en el conjunto de prueba.

  La variable de clasificación es nominal y puede tomar los siguientes dos valores: *tipo-g, tipo-b*. Todas las variables tienen un referencial común, para las cuales se construye un mismo dominio constituido por siete etiquetas.

- **Base de datos SOYBEAN**

  Descripción: Esta base de datos se tomó como punto de partida para el desarrollo de un sistema experto para la diagnosis de enfermedades de las habas [111]. El conjunto de ejemplos contiene 35 variables predictivas nominales o discretas y una variable de clasificación con 19 clases. Hay 316 ejemplos que contienen valores perdidos para muchas de las variables predictivas que se dividen en una única partición en fichero de entrenamiento y prueba.