Instituto Superior Politécnico "José A. Echeverría"

Facultad de Ingeniería Industrial

Algoritmo para la selección de atributos en la predicción del fracaso empresarial

TESIS PARA OPTAR POR EL TÍTULO DE:

MÁSTER EN TECNOLOGÍAS DE APOYO A LA TOMA DE DECISIONES

Autora:

LIC. Lytyet Fernández Capestany

Tutores:

Dr. Silvio Cálves Hernández

MSc. Bolivar Ernesto Medrano Broche

Asesor:

MSc. Yunier Emilio Tejeda Rodríguez

La Habana, 2014

Dedicatoria

A, mi familia

a ti, B.E.

Agradecimientos

Agradezco a:

- A mi tutor Dr. Silvio Cálves, por todos los consejos y conocimiento trasmitido.
- A mi tutor y compañero MSc. Bolivar E. Medrano por toda la atención, por las horas de trabajo dedicadas, por toda la paciencia que ha tenido conmigo en todos estos años.
- A mi asesor MSc. Yunier E. Tejeda por toda su ayuda y su contribución en el desarrollo inicial de las implementaciones necesarias para la investigación, y sus revisiones desde la lejana Angola.
- A los profesores de la maestría que han contribuido a mi formación.
- A , Hugo por todas las sugerencias y sus aclaraciones oportunas en el curso de esta investigación.
- A mi mamá por su confianza, a mis dos tías Mercy y Niurka por su apoyo incondicional.
- A mis amistades Cea y Yady por estar presente en todos los momentos, Cea gracias por las sugerencias en la revisión del documento.
- A los compañeros de trabajo que me permitieron tener el tiempo necesario para dedicarlo al desarrollo de esta investigación.

Resumen

En la predicción del fracaso empresarial resulta útil contar con técnicas que permitan seleccionar las razones financieras que tengan alto poder predictivo. En este trabajo se propone un algoritmo que en cada iteración selecciona un subconjunto de atributos por medio de la descomposición matricial aleatoria. Cada subconjunto de atributos es evaluado por su capacidad predictiva que se calcula mediante un clasificador (k-Vecinos más Cercanos y Análisis Discriminante Lineal) a partir de la Validación Cruzada (Leave One Out y Bootstrap). En el algoritmo la evaluación de los atributos mediante el clasificador se puede realizar directamente con los atributos originales o con componentes calculadas mediante Mínimos Cuadrados Parciales o Análisis de Componentes Principales. El desempeño del algoritmo se prueba en dos conjuntos de datos que contienen razones financieras de empresas fracasadas y empresas no fracasadas. Los resultados obtenidos son comparables con los reportados en la literatura para los mismos conjuntos de datos.

Abstract

In the prediction of business failure is useful to have techniques to select financial ratios that have high predictive power. In this paper an algorithm in each iteration selects a subset of attributes by random matrix decomposition is proposed. Each data set of the attributes is evaluated for its predictive capacity is calculated by a classifier (k-Nearest Neighbors and Linear Discriminant Analysis) by Cross Validation (Leave One Out and Bootstrap). In the algorithm the evaluation of attributes by using the classifier can be done directly with the original attributes or can be used the components get by Partial Least Squares or Principal Component Analysis. The performance of the algorithm is tested on two datasets containing financial ratios of default and non-default enterprises. The results are comparable with those reported in the literature for the same data sets.

Índice general

Índice de algoritmos

Índice de figuras

Índice de tablas

0.1. Introducción

Ahora que hemos reunido tantos datos, qué hacemos con ellos?

U.M. Fayyad.

La actual crisis económica global está teniendo una fuerte incidencia en la actividad empresarial, provocando la desaparición de miles de empresas. *"Las consecuencias del fracaso empresarial no se limitan a las personas, empresas u organizaciones que han establecido de forma directa una relación con la empresa fallida. A menudo estas consecuencias se extienden al entorno empresarial y, por ende, al conjunto de la economía y al entorno social de un país o de una región. Por ejemplo, los países o regiones menos desarrollados son, generalmente, muy sensibles a los fracasos de sus empresas, especialmente, si éste afecta a alguna de las empresas con un importante impacto en la economía regional o nacional"* Alfaro (2006).

En Cuba no existe una ley que considere explícitamente la quiebra empresarial. Aunque ya en el lineamiento 17, se hace referencia a que: *"las empresas que obtengan resultados negativos o muestren sostenidamente en sus balances financieros pérdidas, serán sometidas a un proceso de liquidación o se podrán transformar en otras formas de gestión no estatal (...)"* PCC (2011). La apertura y la actualización que experimenta nuestra economía nos ha motivado a adentrarnos en esta área del conocimiento.

La toma de decisiones basada en la información económico-financiera es un proceso que tiene incidencia directa en la supervivencia de las empresas actuales. Las repercusiones socioeconómicas que lleva asociadas el fracaso empresarial, han generado un interés creciente en el ámbito académico. En este sentido, múltiples investigaciones van encaminadas a encontrar indicadores que permitan detectar las posibles situaciones de crisis, de forma que se puedan tomar medidas correctivas que eviten el fracaso financiero y sus consecuencias Casanova (2011).

Las razones financieras son expresiones matemáticas de la relación entre dos actividades contables. Su estudio permite conocer la evolución de los valores a través del tiempo, y pueden ser tomadas como una señal de alarma de algún tipo de situación potencialmente peligrosa para las empresas Sanchis (1999). También muestran las condiciones en que opera la empresa con respecto al nivel

de liquidez, solvencia, endeudamiento, eficiencia, rendimiento y rentabilidad, facilitando la toma de decisiones gerenciales, económicas y financieras en la actividad empresarial Nava (2009); Alarcón y Ulloa (2012). Las razones financieras a pesar de ser un instrumento de uso frecuente, cuyo buen diseño y conocimiento permiten resolver algunos aspectos concretos para la toma de decisiones financieras, tienen una capacidad limitada para cuantificar de forma eficiente el éxito o fracaso empresarial Ibarra (2009).

En la literatura aparecen gran cantidad de trabajos para la predicción del fracaso empresarial, que toman como datos las razones financieras. Los primeros estudios referidos a la estimación de modelos predictivos del fracaso empresarial basados en razones financieras, estuvieron a cargo de Beaver (1966) y Altman (1968) con sus modelos univariables y modelos multivariable, respectivamente Beaver (1966); Altman (1968). En las décadas posteriores aparecen en la literatura gran variedad de técnicas para la clasificación de empresas según su estado financiero. Algunas de las principales técnicas que aparecen reportadas son: el Análisis Discriminante Lineal (LDA) Deakin (1972); Blum (1974); Altman y Eisenheis (1978), la Regresión Logística Ohlson (1980); Lo (1986); Premachandra y otros. (2009), las Redes Neuronales Atiya (2001); Du Jardin (2009), Máquina de Soporte Vectorial (SVM) Kim y Sohn (2010), entre otras.

La reducción de la dimensión de una matriz de datos aparece como un problema a resolver en varias áreas del conocimiento donde se aplica la Minería de Datos (DM). Al llevar a cabo la clasificación es útil reducir la dimensión de los datos, pues mejora el desempeño de los clasificadores. Las técnicas para la reducción de la dimensión se suelen clasificar en dos tipos: las de construcción de atributos y las de selección de atributos van der Maaten y otros. (2009).

Una de las técnicas de construcción de atributos más populares es el Análisis de Componentes Principales (PCA) van der Maaten y otros. (2009). Esta técnica tiene el inconveniente de que rompe con la representación original de las variables, lo cual enrarece la interpretación de los resultados. Una ventaja del *PCA* es que los atributos que construye, tienen la característica de ser ortogonales entre sí. Este rasgo hace que el *PCA* sea muy empleado en condiciones en que existe multicolinealidad. En cambio las técnicas para la reducción de la dimensión basadas en la selección de atributos, tienen a su favor que no rompen con la semántica de los datos, lo que facilita la interpretación de los resultados. Tienen la desventaja de no tratar la multicolinealidad lo que puede limitar el uso de algunos clasificadores.

La Regresión por Mínimos Cuadrados Parciales (PLS) es una técnica de reducción de la dimensión de la matriz de datos basada en la construcción de atributos que usualmente se denominan componentes o variables latentes de Jong (1993). Esta técnica combina las características de la Regresión Lineal y el *PCA*, que puede ser utilizada en condiciones donde el número de observaciones es mucho menor que el número de atributos. Esta técnica al igual que *PCA* construye componentes ortogonales entre sí. *PLS* ha sido aplicado con éxito para reducir la dimensión en problemas de clasificación, en el área de la quimiometría y la bioinformática Wold y otros. (2001); Nguyen y Rocke (2002); Barker y Rayens (2003); Dai y otros. (2006); Boulesteix y Strimmer (2007). Recientemente *PLS* ha sido empleado de manera combinada con *SVM* y con el *LDA*, para la predicción del fracaso empresarial Yang y otros. (2011); Serrano y Gutiérrez (2013).

Otra técnica para la reducción de la dimensión de la matriz de datos, es la descomposición matricial *CUR* Mahoney y Drineas (2009). Esta técnica, al igual que la descomposición en valores singulares (SVD) Golub y Van Loan (1996), permite obtener aproximaciones matriciales de menor rango para una matriz de datos. El acrónimo *CUR* viene dado porque esta técnica consiste en descomponer la matriz de datos $A \approx C \times U \times R$ en tres matrices donde C es un subconjunto de columnas de A , R es un subconjunto de filas de A y U una matriz de mezcla o de transición. Dada su naturaleza, la descomposición matricial *CUR* resulta una herramienta interesante cuya utilización puede ser un auxiliar importante en la interpretación cabal de los resultados de un análisis exploratorio de datos Mahoney y Drineas (2009); Thurau y otros. (2012).

Al realizar una búsqueda no exhaustiva en el GoogleScholar y MicrosoftAcademic, con criterios de búsqueda[1] que relacionan, la predicción de fracaso empresarial y las descomposiciones o factorizaciones matriciales, sólo se encontraron los trabajos Ribeiro y otros. (2009); Chen y otros. (2011). En ellos se aborda el problema de la selección de razones financieras mediante la factorización de matrices no negativa (NMF). Esta técnica es similar a la descomposición matricial *CUR*, pero consiste en descomponer la matriz de datos $A_{m\times n} \approx W_{m\times k} \times H_{k\times n}$, donde W se interpreta como un conjunto de k factores, que cada uno tiene asociado un peso por cada variable de los datos originales. Tras hacer un análisis bibliométrico y no encontrar una cantidad suficiente de trabajos se puede afirmar que la aplicación de la descomposición *CUR* en el contexto de la predicción del fracaso empresarial, resulta un tema de investigación interesante y novedoso.

[1] Matrix factorization + bankruptcy, Matrix descomposition + bankruptcy or financial distress or financial default

La gran cantidad de razones financieras presentes en la literatura contable y financiera ha incidido en que los investigadores se refieran a una misma razón con diferentes nombres, o se refieran con un mismo nombre para un conjunto de razones Ibarra (2001); Tascón y Castaño (2010, 2012). En la predicción del fracaso empresarial, usualmente se tiene como fuente de datos una elevada cantidad de razones financieras, que se han calculado para cada una de las organizaciones en estudio. Esta cuestión provoca que los análisis, a menudo, se realicen con información duplicada e irrelevante.

Lo anterior pone de manifiesto la necesidad de contar con la capacidad de seleccionar aquellas razones que resulten más importantes para predecir el fracaso empresarial Tsai (2009). Por esto, se hace necesario aplicar técnicas que al reducir la dimensión de los datos no cambien su estructura original, para luego conformar modelos que puedan ser comprensibles por expertos en análisis financieros.

Por esta razón se plantea el siguiente ***problema de investigación***: Cómo reducir la dimensión de los datos en problemas de aprendizaje supervisado en el contexto de la predicción del fracaso empresarial, sin alterar la estructura de los atributos originales.

Para darle solución al problema científico señalado anteriormente, se plantea como ***hipótesis***: Si se emplea la descomposición CUR de manera combinada con un clasificador y un método de obtención de componentes, se podrá diseñar un algoritmo eficiente, que permita la obtención de subconjuntos de atributos con alto poder predictivo en la clasificación de empresas según su situación financiera.

El ***objetivo general*** de esta investigación consiste en diseñar un algoritmo de selección de atributos, para ser usado en la predicción del fracaso empresarial.

Para alcanzar la meta propuesta se identificaron los objetivos ***específicos*** que se listan a continuación:

1. Identificar los elementos de los algoritmos de selección de atributos que aparecen en la literatura para tomarlos como base en el desarrollo de la propuesta.
2. Diseñar un algoritmo para la selección de atributos mediante el empleo combinado de la descomposición matricial CUR, un método de construcción de componentes y un clasificador.
3. Probar el algoritmo propuesto con el empleo de conjuntos de datos que aparecen en la literatura.

Este trabajo consta de tres capítulos. En el capítulo 1 se recogen los aspectos teóricos de carácter general tenidos en cuenta en el diseño del algoritmo de selección de atributos. En el capítulo 2 se trata todo lo referido al diseño y la implementación del algoritmo que se propone para la selección de

razones financieras. En el capítulo 3 se muestran los resultados que alcanzan las diferentes variantes del algoritmo diseñado en dos conjuntos de datos usados en la literatura. Finalmente se incluyen las conclusiones y las líneas para el trabajo futuro. Este trabajo también cuenta con varios anexos donde se incluyen gráficas y algoritmos que completan la explicitación de la investigación.

1. Fundamentos teóricos

En este capítulo se abordan algunos aspectos de carácter general relacionados con el problema en estudio. Se exponen los elementos básicos del problema de clasificación y la selección de atributos. Por esta razón se incluyen conceptos y definiciones que ayudan a comprender el resto de la investigación. Además se describen de manera general los clasificadores, los métodos de obtención de componentes y los métodos de validación de la clasificación que se emplean en el diseño del algoritmo.

1.1. Extracción de conocimiento

La Minería de Datos (*DM*) debe su nombre a las similitudes encontradas entre buscar valiosa información de negocio en grandes bases de datos y minar una montaña para encontrar una veta de metales valiosos. Su tarea fundamental es la de encontrar modelos inteligibles a partir de los datos, y para que el proceso sea efectivo debe ser automático, generando patrones que ayuden en la toma de decisiones beneficiosas. Hoy la *DM* está compuesta por un conjunto de métodos matemáticos y técnicas de software para el análisis inteligente de datos, la búsqueda de regularidades y tendencias en los mismos, aplicados de forma iterativa e interactiva.

En la *DM* para obtener conclusiones válidas y útiles es necesario complementar este proceso con una adecuada preparación de los datos, previa al proceso de minería y un análisis posterior de los resultados obtenidos. Por esta razón varios autores enmarcan la *DM* en un proceso más amplio que es denominado, Descubrimiento de Conocimiento en Bases de Datos (*KDD*). En la figura (1.1.1) se muestra un esquema de este proceso. Esta investigación se centra principalmente en la segunda etapa del proceso del *KDD*, es decir en la preparación de datos, y más concretamente en la selección de características.

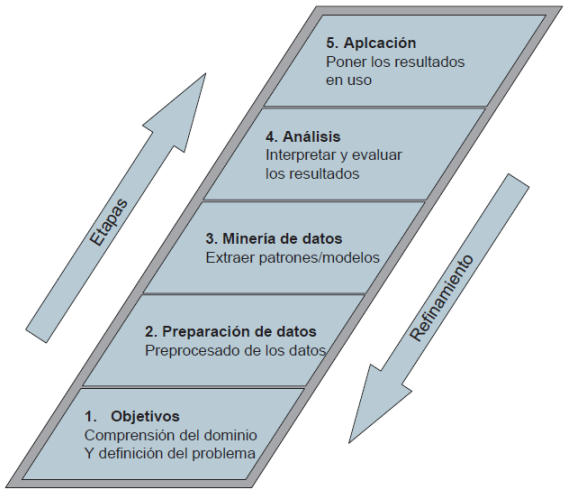


Figura 1.1.1.: Proceso de extracción de conocimiento Sánchez (2005).

1.2. Representación de los datos

A continuación se establecen algunas definiciones que describen formalmente los conceptos que se manejarán a lo largo de este documento, tomando como referencia las definiciones propuestas en Sánchez (2005).

Definición 1. (Dominio) Se le denomina a un conjunto de valores del mismo tipo. Aunque existen distintas clasificaciones de los dominios, para los propósitos de este trabajo se considera el dominio de tipo continuo (conjunto infinito de valores reales).

La Contabilidad se encarga de registrar, clasificar y resumir en términos monetarios, las operaciones y transacciones que ocurren en los diferentes procesos económicos Brigham y Houston (2009). En general los datos que se manejan pertenecen a un dominio de valores reales.

Definición 2. (Universo de discurso) Es el entorno donde se define un determinado problema y viene representado como el producto cartesiano de un conjunto finito de dominios.

El universo de discurso al que se le presta atención en este trabajo es al análisis de los estados financieros de las organizaciones. Los estados financieros son el producto final del proceso contable y la entrada al comienzo de una etapa analítica.

1.2.1. Atributos

Otra definición indispensable que se debe tener en cuenta en la representación de los datos es la de atributo. Según Sánchez (2005) atributo se define como:

Definición 3. (Atributo) Un atributo, o también denominado característica, es la descripción de alguna medida existente en el universo de discurso que toma valores en un determinado dominio. También en el contexto de este trabajo se emplea el término de variable como sinónimo de atributo y se denota como x_j. En este trabajo se considera como atributos a las razones financieras.

1.2.1.1. Razones financieras

Las razones financieras son los tipos de atributos más utilizados históricamente para la predicción del fracaso empresarial al usar el aprendizaje supervisado. Ayudan a estandarizar la información de acuerdo al tamaño de la empresa, proceso que hace posible, completar la tarea de clasificación o de comparación que intervienen en cualquier proceso de predicción y comparar el efecto por el tamaño Du Jardin (2009). Desde el punto de vista financiero se define razón financiera de la manera siguiente:

Definición 4. (Razón financiera) Es la expresión de la relación entre actividades contables, que tienen valores propios o absolutos y presentan un nivel o cantidad determinada de actividad de la empresa Weston y Brigham (1994).

Las razones financieras permiten conocer los resultados obtenidos en cada ejercicio económico de la organización, es decir, cuanto se ha ganado o perdido en un período de tiempo determinado, las causas y conexiones entre los hechos y fenómenos monetarios que ocurren en el ámbito interno de la empresa Xiomara (2005). Pocas son las cifras en un estado financiero que pueden considerarse altamente significativas por sí mismas, lo importante es su relación con otras cantidades Demestre y otros. (2005).

1.2.2. Clases

Definición 5. (Clase) Es un atributo especial de salida, que indica la pertenencia a un determinado grupo de casos. Se denomina etiquetas de clase al conjunto o dominio de valores que la clase puede

tomar (nominal en el caso de la clasificación). La clase es el atributo sobre el cual se realiza la predicción, por lo que es también denominada atributo de decisión, para diferenciarla del resto de los atributos denominados de condición. El atributo clase se representa usualmente con la letra Y, teniendo k valores posibles $y_1, ..., y_k$.

Los valores de las etiquetas de clases que se emplearán en esta investigación tendrán los valores siguientes: $y \in \{1, 2\}$.

1.2.3. Conjunto de datos

Definición 6. (Caso, instancia, observación) Un ejemplo, muestra, instancia o caso es una tupla del universo de discurso representada por un conjunto de valores de atributos de un dominio determinado, y una etiqueta de clase que lo clasifica. Un caso, observación se denota como e_j.

Las observaciones que son consideradas en esta investigación corresponden a empresas que pueden estar etiquetadas como ($y = 1$) solvente (sana o no quebrada) o ($y = 2$) insolvente (no sana o quebrada) .

Definición 7. (Conjunto de datos) Es un subconjunto finito de observaciones e_j, donde $j = 1, ..., m$. Un conjunto de datos, o base de datos, se caracteriza por el número de observaciones m que contiene, por el número n de atributos y el tipo de estos. Los conjuntos de datos son denotados como D.

En el análisis financiero las cantidades de datos contenidos en los estados pueden sobrepasar los 120 conceptos y al utilizar la técnica de razones financieras, puede crecer esta cifra exponencialmente. Esto deriva en una dificultad en el manejo de una cantidad de datos para llevar a cabo el análisis, además de que la información puede ser similar y redundante García y otros. (2012). El número de razones que se analicen debe ser reducido pues un número excesivo de estas requeriría mucho tiempo, para obtener una apreciación global al no poder relacionarlas entre sí Sanchis (1999).

1.3. Clasificación

Cuando se habla de clasificación puede tener dos significados distintos, se puede tener un conjunto de observaciones con el objetivo de establecer la existencia de clases o grupos en los datos. O se

puede saber que existen determinadas clases, y que el objetivo sea establecer una regla por la que se llegue a clasificar una nueva observación dentro de una de las clases existentes. El primer tipo se conoce como aprendizaje no supervisado (o clustering), el segundo como aprendizaje supervisado Hastie y otros. (2001). Este trabajo presta mayor atención al aprendizaje supervisado.

1.3.1. Clasificadores

En la estadística tradicionalmente se ha utilizado para este propósito el *LDA*, pero en los últimos tiempos se han desarrollado nuevas técnicas, en parte gracias al desarrollo experimentado en la capacidad de cómputo. A continuación se incluye una descripción general del funcionamiento de los clasificadores considerados en este trabajo.

1.3.1.1. Análisis Discriminante Lineal

El Análisis Discriminante Lineal (LDA) es uno de los procedimientos de clasificación más antiguos y es el que mayoritariamente se utiliza en los paquetes estadísticos. La idea que sigue es la de dividir el espacio muestral mediante una serie de líneas en el espacio bidimensional, planos en el caso de tres dimensiones, y en general, hiperplanos para muchas dimensiones. El *LDA* se puede considerar como un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada una de las clases, y las variables independientes (atributos), son continuas y determinan a qué clase pertenecen las observaciones.

Tiene como un primer objetivo encontrar relaciones lineales entre las variables continuas que mejor discriminen las clases, dadas las observaciones. Un segundo objetivo es construir una regla de decisión que asigne una observación nueva, que no se sabe clasificar previamente, a una de las clases prefijadas con un cierto grado de riesgo.

Al aplicar el *LDA* es necesario considerar una serie de supuestos:

- Existencia de una variable categórica Y y un conjunto X de variables de intervalo o de razón que son independientes respecto Y.
- Es necesario que existan al menos dos grupos, y para cada grupo se necesitan dos o más casos.
- El número de variables discriminantes debe ser menor que el número de objetos menos 2 : $x_1, \ldots, x_p$; donde $p < (n-2)$ y n es el número de objetos.

- Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.
- El número máximo de funciones discriminantes es igual al mínimo entre el número de variables y el número de grupos menos 1 (con q grupos, $(q-1)$ funciones discriminantes).
- Las matrices de covarianzas dentro de cada grupo deben ser aproximadamente iguales.
- Las variables continuas deben seguir una distribución normal multivariable.

Los fundamentos matemáticos del *LDA* pueden ser consultados en Hair y otros. (1999); Johnson y Castellanos (2000). El pionero en abordar la predicción del fracaso empresarial basado en razones financieras empleando el *LDA*, fue Altman, con sus modelos multivariable Altman (1968); Altman y Eisenheis (1978).

1.3.1.2. K Vecinos más Cercanos

La técnica de Vecinos más Cercanos (NN) basan su criterio de aprendizaje en la hipótesis de que los miembros de una población suelen compartir propiedades y características con los individuos que los rodean, de modo que es posible obtener información descriptiva de un individuo mediante la observación de sus vecinos más cercanos. Esta técnica fue enunciada formalmente por Cover y Hart en Cover y Hart (1967). Desde entonces, este algoritmo se ha convertido en uno de los métodos de clasificación más usados Dasarathy (1991); Denoeux (1995); Kleinberg (1997); Archana y Elangovan (2014). La regla de clasificación *NN* se resume básicamente en el siguiente enunciado: Sea $D = \{e_1, ..., e_m\}$ un conjunto de datos con m observaciones etiquetadas, donde cada una contiene n atributos $(x_{j1}, ..., x_{jn})$, pertenecientes al espacio métrico X, y una clase $y_l \in \{y_1, ..., y_k\}$. La clasificación de una nueva observación $e*$ cumple que:

$$e* \dashv y_l \text{ si solo sí para todo } j \neq i \text{ se tiene que } d(e*, e_i) < d(e*, e_j)$$

donde $e* \dashv y_l$ indica que la asignación de la etiqueta de clase y_l a la observación $e*$, y d expresa una métrica definida en el espacio métrico X.

Así, un ejemplo es etiquetado con la clase de su vecino más cercano según la métrica definida por la distancia d. La elección de esta métrica es primordial, ya que determina qué significa más cercano.

La aplicación de métricas distintas sobre un mismo conjunto de entrenamiento puede producir resultados diferentes. Sin embargo, no existe una definición previa que indique si una métrica es buena o no. Esto implica que es el experto quien debe seleccionar la medida de distancia más adecuada. La regla *NN* puede generalizarse calculando los k Vecinos más Cercanos y asignando la clase mayoritaria entre esos vecinos, tal generalización se denomina (kNN). Este algoritmo necesita la especificación a priori de k, que determina el número de vecinos que se tendrán en cuenta para la predicción. Al igual que la métrica, la selección de un k adecuado es un aspecto determinante. La elección de un k adecuado y una métrica conveniente, ha sido tratado en Dudani (1976); Wettschereck y Dietterich (1995).

1.3.2. Evaluación de la ejecución de un clasificador

El objetivo de cualquier sistema de clasificación entrenado en un conjunto de ejemplos, es aprender a clasificar correctamente las nuevas observaciones que se le presenten. Para evaluar el comportamiento futuro del clasificador ante nuevos ejemplos, se intentará estimar la precisión del clasificador, al mismo tiempo esta estimación ayuda a seleccionar el sistema de clasificación más apropiado en cada caso para el problema concreto (número y tipo de atributos) Alfaro (2006).

La forma más habitual de medir la eficiencia de un clasificador es la precisión predictiva. Cada vez que se presenta un nuevo caso, el clasificador, debe tomar una decisión sobre la etiqueta que le va a asignar. Considerando un error como una clasificación incorrecta de una observación, se puede calcular fácilmente la tasa de error, o su complemento, la tasa de clasificación correcta.

Definición 8. (Precisión) Se denomina precisión de un clasificador, al resultado de dividir el número de clasificaciones correctas por el número total de muestras.

$$P = \frac{CC}{N} \tag{1.3.1}$$

Donde CC es la cantidad de individuos clasificados correctamente y N es la cantidad de observaciones del conjunto de datos. La precisión es una buena estimación de cómo se va a comportar el modelo para datos desconocidos similares al conjunto de datos de prueba. Sin embargo, si se calcula la precisión sobre el propio conjunto de datos utilizado para generar el modelo, se obtiene con

frecuencia una precisión mayor a la real, es decir, serán estimaciones muy optimistas por utilizar las mismas observaciones en la inducción del algoritmo y en su comprobación Hair y otros. (1999). Esta separación es necesaria para garantizar la independencia de la medida de precisión resultante, de no ser así, la precisión del modelo será sobrestimada Hastie y otros. (2001). Para tener seguridad de que las predicciones sean robustas y precisas, se consideran dos etapas en el proceso de construcción de un modelo, entrenamiento y prueba, partiendo los datos en dos conjuntos.

Para evaluar un clasificador cuando las clases no están desvalancedas, se suele usar la tasa promedio de clasificación correcta de cada clase (expresión) Ferri y otros. (2009):

$$PM = \frac{1}{c}\sum_{i=1}^{c}\frac{CC_i}{C_i} \tag{1.3.2}$$

Donde CC_i es la cantidad de observaciones de la clase i clasificadas correctamente, C_i es la cantidad de observaciones de la clase i y c es la cantidad de clases. Para evitar el sobreajuste de un clasificador resulta necesario estimar el modelo con una porción de los datos (conjunto de entrenamiento) y luego comprobar su validez con el resto de los datos (conjunto de prueba), dependiendo de como se haga la partición del conjunto y el tamaño de la muestra será la técnica de evaluación a emplear.

1.3.2.1. Validación Cruzada

La Validación Cruzada consiste en dividir aleatoriamente el conjunto original de datos en k subconjuntos mutuamente excluyentes con aproximadamente el mismo tamaño. Cada uno de estos subconjuntos se utiliza como conjunto de prueba para el clasificador construido a partir de los $k-1$ subconjuntos restantes, de tal forma que al final se tendrán k clasificadores distintos con sus respectivas tasas de error en el conjunto de prueba. La tasa de error estimada por validación cruzada será la media de las k tasas de error calculadas, ponderando por los tamaños de los subconjuntos si son de distinto tamaño. En el caso de que el tamaño de todos los subconjuntos coincida, no será necesario ponderar por sus tamaños.

Este estimador tiene un sesgo bastante optimista respecto a la tasa de error real. Existen resultados empíricos que muestran que si k es menor que cinco, las estimaciones son demasiado sesgadas Kohavi (1995). En cambio si k está próximo a diez habrá un sesgo aceptable y si k es mayor que 20 las estimaciones serán casi insesgadas. El valor de k (número de particiones) que más se utiliza es 10,

consiguiendo así que en cada prueba la reducción del conjunto de entrenamiento no sea demasiado grande. La desventaja de utilizar este método es el coste computacional, porque repite k veces el proceso de aprendizaje.

1.3.2.2. Leaving One Out

Este método es un caso particular de la Validación Cruzada, donde el conjunto original se divide de forma aleatoria en m subconjuntos de tal forma que cada uno de ellos sólo contiene una observación. Para un determinado método de clasificación y un conjunto de m observaciones, se genera un clasificador utilizando $m-1$ observaciones y se prueba con la observación que se ha dejado fuera (de ahí el nombre de Leaving One Out (LOO), ya que en cada iteración se deja una observación fuera del entrenamiento). Este procedimiento es aún más costoso que la Validación Cruzada con 10 subconjuntos, ya que en este caso hay que construir m clasificadores, cada uno de ellos a partir de $m-1$ observaciones. Por esta razón este método sólo es aconsejable para problemas con un conjunto de datos pequeño. El estimador LOO es casi insesgado, pero su varianza es grande para pequeñas muestras (30 observaciones o menos). Por esta razón para pocas observaciones se recomienda Bootstrap Braga-Neto y Dougherty (2004).

1.3.2.3. Bootstrap

Para datos con un número pequeño de muestras la Validación Cruzada presenta estimadores con una alta varianza, lo que hace que estos sean no fiables Braga-Neto y Dougherty (2004). En estos casos es recomendable usar el método de Bootstrap. La idea del método Bootstrap consiste en tratar los datos muestrales como si constituyesen los datos de toda la población, es decir se utilizan como el universo del que se extraerán muestras con reemplazo Efron y otros. (1979). Este método parte de un conjunto con m observaciones del cual se seleccionan m observaciones, mediante el muestreo con reemplazo para formar el conjunto de entrenamiento. Lo que trae consigo que una observación pueda estar repetida dos o más veces en un mismo conjunto de entrenamiento. Como conjunto de prueba se utilizan aquellas observaciones que no han sido incluidas ninguna vez en el conjunto de entrenamiento. La proporción esperada de ejemplos que pasan a formar parte del conjunto de entrenamiento es de 63,2 % y la proporción esperada de ejemplos en el conjunto de prueba es de

36,8 %. La tasa de error estimada es la media de las tasas de error después de varias iteraciones. Alrededor de doscientas iteraciones se consideran necesarias, para que esta estimación sea buena. Con Bootstrap se intenta para reducir la alta variabilidad que exhibe la Validación Cruzada y *LOO* en muestras pequeñas, lograr un aumento de la eficiencia comparable a un aumento en el tamaño de la muestra en un 60 % Efron y Tibshirani (1994).

1.4. Reducción de la dimensión

En la actualidad en varias áreas del conocimiento existe la necesidad de extraer información a grandes volúmenes de datos. La elevada dimensión de un conjunto de datos afecta el desempeño de los algoritmos existentes para la clasificación, así como la interpretación de estos van der Maaten y otros. (2009). Existen dos aproximaciones para realizar una reducción de la dimensión de un conjunto de datos. Una, es la transformación o construcción de atributos y la otra es la selección de atributos, ambas son técnicas de preprocesado que se usan frecuentemente en la *DM*.

1.4.1. Construcción de atributos

De manera general las técnicas de construcción de atributos consisten en la creación de un nuevo conjunto de atributos a partir de los originales. Los nuevos atributos muchas veces llamados componentes por lo general son ortogonales entre sí, no correlacionados. Existen gran cantidad de técnicas que construyen nuevos atributos estableciendo relaciones lineales o no-lineales con los atributos originales, las cuales han sido empleadas en determinados contextos van der Maaten y otros. (2009). En este trabajo solamente es de interés el Análisis de Componentes Principales y la Regresión por Mínimos Cuadrados Parciales.

1.4.1.1. Análisis de Componentes Principales

El Análisis de Componentes Principales (*PCA*) es una técnica cuyo objetivo principal es hallar combinaciones lineales de los atributos originales de un conjunto de datos, con la propiedad de que exhiban varianza mínima y que a la vez no estén correlacionadas entre sí Mardia y otros. (1979). Para obtener tales combinaciones es necesario construir la matriz de varianzas y covarianzas de los

atributos. El *PCA* permite reducir la dimensionalidad de los datos, ya que transforma el conjunto de los p atributos originales en otro conjunto de q nuevas variables no correlacionadas llamadas componentes principales. De esta manera se puede elegir una cantidad de componentes $q \leq p$ que pueden explicar gran parte de la variabilidad de los datos.

En la primera componente, mientras mayor sea su varianza, mayor será la cantidad de información que se explica en dicha componente. Por ello las sucesivas componentes se ordenan en forma descendente de acuerdo a la proporción de la varianza que resuman cada una de ellas. La primera componente es por lo tanto, la combinación de máxima varianza, la segunda es otra combinación de variables originarias que obedece a la restricción de ser ortogonal a la primera y de máxima varianza, la tercera componente es aún otra combinación de máxima varianza, con la propiedad de ser ortogonal a las dos primeras; y así sucesivamente.

Cuando las variables están correlacionadas en alto grado, las primeras componentes explican una alta proporción de la varianza total, por eso las componentes principales pueden sustituir a los múltiples atributos originales. Esto permite resumir en pocas componentes no correlacionadas gran parte de la información del conjunto de datos.

Sin embargo, las primeras componentes que se supone que explican una alta variabilidad no necesariamente mejoran la predicción cuando se usa en regresión o en clasificación Vega-Vilca (2004). En la clasificación no supervisada, trabajos como el de Yeung y Ruzzo (2001), demuestran que el uso de componentes principales en vez de las variables predictoras originales, no necesariamente mejora y en muchos casos degrada la calidad esperada de clasificación, ellos llegan al extremo de no recomendar su uso. A pesar de lo anterior en este trabajo se emplea el *PCA* para construir componentes, esta decisión se debe a que los conjuntos de datos formados por razones financieras tienen elevada multicolinealidad.

1.4.1.2. Regresión por Mínimos Cuadrados Parciales

La regresión por Mínimos Cuadrados Parciales (*PLS*), fue introducida por Wold (1975), para ser aplicada en las ciencias económicas y sociales, o en el área de la quimiometría años más tarde. El empleo de esta técnica se ha incrementado en los últimos años ya sea por su versatilidad o por los resultados alcanzados en disímiles aplicaciones Vega-Vilca y Guzmán (2011). Con *PLS* se generaliza

y se combinan las características del *PCA* y el Análisis de Regresión Múltiple.

La regresión *PLS* consta de dos pasos fundamentales. En el primero se transforma la matriz de datos $X_{m\times n}$ con ayuda del vector de clases $Y_{m\times 1}$ en una nueva matriz $T_{m\times n}$ formada por atributos comúnmente denominados variables latentes o las componentes *PLS*. Esto contrasta con el *PCA* en el cual las componentes son obtenidas usando sólo la matriz de atributos X. Las componentes $T_{m\times n}$ conservan la relación de dependencia con el vector de clases, esto le confiere cierta ventaja sobre *PCA*. Un segundo paso de *PLS* consiste en calcular el modelo de regresión estimado usando el vector de clases original y las componentes *PLS*. La reducción de la dimensión puede ser aplicada directamente sobre las componentes ya que estas son ortogonales entre sí. El número de componentes necesarias para el análisis de regresión debe ser mucho menor que el número de atributos originales.

La idea de la regresión *PLS* es maximizar el cuadrado de la covarianza entre el componente $T_h = Xw$, y la variable respuesta Y, sujeto a $w'w = 1$. El componente T_h está definido como una combinación lineal de las predictoras, tal que $w \neq 0$. Existen varios algoritmos para *PLS*, los más populares son el NIPALS y el SIMPLS propuestos en Geladi y Kowalski (1986) y de Jong (1993) respectivamente. El algoritmo NIPALS (algoritmo (1.1)) es menos eficiente que el algoritmo SIMPLS, pero tiene una estructura más simple y comprensible. Por esta razón a continuación se incluye el NIPALS en lugar del SIMPLS, a pesar de que su homólogo sea más usado en la actualidad. En el anexo (A) se muestra la estructura del SIMPLS.

Algoritmo 1.1 Algoritmo NIPALS Trygg y Wold (2002).

1: *Entrada:* $X(h)$, Y, $h = 0$
2: *Para* $h = 1$ *hasta* n
3: $w = cov(Y,X)$: *normalizar* w
4: $T_h = Xw$
5: $v = \left(T_h'Y\right) / \left(T_h'T_h\right)$
6: $b = \left(T_h'X\right) / \left(T_h'T_h\right)$
7: $X(h) = X(h-1) - T_h b$
8: $Y(h) = Y(h-1) - T_h v$
9: *Próximo* h

A continuación se describen los principales pasos de este algoritmo:

1. El algoritmo NIPALS para la regresión *PLS* tiene como entrada el conjunto de datos $X_{m\times n}$

estandarizado por columnas y el vector de clases $Y_{m\times 1}$. El conjunto de datos puede ser escrito como $X = (X_1, \ldots, X_n)$, donde $X_1, \ldots, X_n$ son los atributos.

2. Ciclo para el cálculo de las componentes *PLS*.
3. Se calcula el vector $w = (w_1, \ldots, w_p)'$, donde cada elemento w_i es la covarianza del vector de clases y el atributo X_i. Finalmente w es normalizado a la unidad.
4. Se calcula la componente *PLS* de la manera siguiente, $T_h = Xw = (X_1, \ldots, X_p) \cdot (w_1, \ldots, w_p)'$.
5. Se calcula el coeficiente de regresión simple del atributo sobre la componente T_h, calculado en el paso anterior.
6. Se calcula el vector $b = (b_1, \ldots, b_p)$, cada elemento de este vector es el coeficiente de regresión simple de X_i sobre T_h.
7. Se actualiza el conjunto de datos $X(h)$.
8. Se actualiza el vector de clases $Y(h)$.
9. Se procede al cálculo de la próxima componente *PLS*, a partir del paso 3.

1.5. Selección de atributos

Al usar *PCA* para reducir la dimensión de un determinado conjunto de datos se construyen nuevas componentes que son combinaciones lineales de los atributos de la matriz de datos originales. Como las primeras componentes son las que resumen la mayor cantidad de la variabilidad de los datos, se toman para realizar los respectivos análisis. Este proceder resulta conveniente cuando la matriz tiene entre 20 y 30 columnas ya que quizás de 3 a 4 componentes resumen una cantidad considerable de la variabilidad de los datos. Al aumentar la cantidad de atributos de las observaciones, el análisis a partir del *PCA* se enrarece pues el número de componentes que resumen cierto nivel de variabilidad, también aumenta. Entonces la interpretación de los vectores propios con vistas a la formación de grupos, o en general, a la clasificación de los individuos según su disposición en el espacio generado por las primeras componentes principales se hace difícil. Esta es una de las razones de mayor peso que se tienen en cuenta a la hora de aplicar técnicas de selección de atributos, para la reducción de la dimensión. Las técnicas de selección de atributos tienen entre sus principales ventajas la preservación de la semántica de los datos, lo que facilita la interpretación de los resultados por parte de

expertos del campo del conocimiento de donde proceden los datos Saeys y otros. (2007).

1.5.1. Proceso de selección

En la selección de atributos se intenta escoger el subconjunto mínimo de atributos de acuerdo con dos criterios: que la tasa de aciertos no descienda significativamente y que la distribución de la clase resultante, sea lo más semejante posible a la distribución de la clase original dados todos los atributos Sánchez (2005).

En resumen, el resultado de la selección de atributos sería:

- Menos atributos: los algoritmos pueden aprender más rápido.
- Mayor exactitud: el clasificador generaliza mejor.
- Resultados más simples: más fácil de entender.

La selección de atributos se puede considerar como un problema de búsqueda Siedlecki y Sklansky (1988); Guyon y Elisseeff (2003), en un espacio de estados, donde cada estado corresponde con un subconjunto de atributos, y el espacio engloba todos los posibles subconjuntos que se pueden generar. Un algoritmo de selección de atributos en general, se basa en dos pasos básicos: generación de subconjuntos de atributos y evaluación de subconjuntos. En la generación de nuevos subconjuntos se define un punto de partida, para que siguiendo una estrategia se recorra el espacio de búsqueda hasta que se cumpla un criterio de parada.

1.5.2. Algoritmos para la selección de atributos

En la literatura científica existen abundantes referencias de trabajos sobre la selección de atributos. La clasificación de los algoritmos para la selección de atributos aún no está unificada. Una de las más aceptadas es la clasificación de Langley en Langley (1994), donde se dividen los algoritmos en dos categorías: algoritmos de filtro y algoritmos envolventes (wrapper). Otras clasificaciones importantes son las de Liu y colaboradores Yu y Liu (2004); Liu y Yu (2005), donde sus principales aportes son la agrupación de los tipos de búsquedas de atributos en tres categorías: completa, heurística y aleatoria. También la dirección (hacia adelante y hacia atrás) en que se lleva a cabo la búsqueda en la selección de atributos, suele ser tomada como criterio para clasificar los algoritmos Sánchez

(2005). Existen también los algoritmos híbridos para la selección de atributos los cuales combinan la estrategia de los mencionados anteriormente.

1.5.2.1. Algoritmos de filtro

En los algoritmos de filtro, se elimina o filtran, los atributos irrelevantes antes de que se produzca el aprendizaje. Después de seleccionar los atributos más relevantes se puede aplicar cualquiera de los métodos de clasificación. Los algoritmos de filtro, evalúan los atributos de acuerdo con heurísticas basadas en características generales de los datos e independientes del método de clasificación a aplicar. También pueden verse como métodos de filtrado, aquellos métodos que construyen atributos de orden superior a partir de los originales, los ordenan en función de la varianza que explican y seleccionan los mejores. El *PCA* es el ejemplo más conocido de este tipo Alfaro (2006).

1.5.2.2. Algoritmos envolventes

En los métodos de tipo envolvente Kohavi y John (1997), la selección de atributos y los algoritmos de aprendizaje no son elementos independientes, ya que emplean un clasificador para evaluar la calidad de cada conjunto de atributos seleccionado en cada momento de la búsqueda. El principal argumento para utilizar estos métodos envolventes es que cada método de clasificación tendrá un sesgo en la fase inductiva y, por tanto, para estimar la precisión del subconjunto de atributos a seleccionar es mejor tener en cuenta el sesgo del método de clasificación que se va a usar después y no el sesgo de otra medida distinta de precisión Alfaro (2006). En cambio el principal inconveniente de estos métodos frente al método de filtrado, es el coste computacional tan grande debido al esfuerzo que supone repetir el proceso de aprendizaje para cada subconjunto de atributos considerado.

1.5.2.3. Dirección de búsqueda

Como se mencionó en el apartado (1.5.1) la selección de atributos puede ser vista como un proceso de búsqueda. En este sentido es conveniente aclarar que se entiende por la dirección de búsqueda en la selección de atributos. Se entiende por dirección de búsqueda, la relación entre los atributos de un subconjunto con el siguiente, al realizar el recorrido a través del espacio de búsqueda se puede seguir uno de los siguientes esquemas:

1. Secuencial hacia adelante: este esquema empieza con el conjunto vacío y se añaden secuencialmente atributos del conjunto original. En cada iteración, se elige el mejor atributo entre los no seleccionados, basándose en alguna función de evaluación. El algoritmo puede parar cuando llegue a un número determinado de atributos seleccionados, o cuando la evaluación sobrepase un valor prefijado, o simplemente, al dejar de aumentar la evaluación.
2. Secuencial hacia atrás: se comienza con el conjunto completo de atributos y se eliminan secuencialmente los atributos. En cada iteración, basándose en alguna función de evaluación, se escoge el atributo menos relevante de entre los originales, eliminándolo del conjunto. Al igual que en el caso anterior, el algoritmo puede parar cuando llegue a un número determinado de atributos en el subconjunto, o cuando la evaluación no llegue a un valor prefijado, o simplemente, al dejar de aumentar la evaluación.
3. Aleatoria: no se conoce el punto de inicio, ni cual es el siguiente subconjunto generado. Se pasa de un subconjunto a otro sin ningún orden establecido, añadiendo o eliminando uno o varios atributos. De esta forma se pretende evitar la elección de subconjuntos de atributos que sean óptimos locales, como puede ocurrir en los casos anteriores. El criterio de parada suele ser la ejecución de un cierto número de iteraciones.

Otros esquemas se pueden obtener variando o mezclando algunos de los anteriores, como el bidireccional, que realiza la búsqueda hacia adelante y hacia atrás al mismo tiempo.

1.5.2.4. Estrategias de búsquedas

Para un conjunto de datos con n atributos, existen 2^n subconjuntos candidatos. Una búsqueda exhaustiva en este espacio es totalmente ineficiente, incluso para un n pequeño, siendo necesario el uso de diferentes estrategias para abordar este problema. A continuación, se explican brevemente las características más importantes de los tipos de estrategias de búsqueda completa, secuencial y aleatoria según Liu y colaboradores Liu y Yu (2005):

1. Completa: esta búsqueda garantiza la localización de un resultado óptimo conforme a un criterio dado. Si para seleccionar los subconjuntos óptimos se ha de examinar todos los conjuntos posibles del espacio, coincidirá con la exhaustiva. Sin embargo, según la medida de evaluación utilizada, puede no ser necesario examinar todos los subconjuntos posibles. Se puede

decir que una búsqueda exhaustiva siempre es completa, pero a la inversa no se cumple en todos los casos.

2. Secuencial: son estrategias que realizan una búsqueda parcial a través del espacio formado por los atributos, con el riesgo de no encontrar subconjuntos óptimos. Lógicamente son algoritmos más rápidos que los pertenecientes al grupo anterior. En este grupo se encuentran distintas variantes de la técnica greedy, que añaden o eliminan atributos uno a uno.
3. Aleatoria: este tipo de estrategias, al contrario de las anteriores, busca a través del espacio formado por los atributos de manera aleatoria, es decir, no existe un estado siguiente o anterior que se pueda determinar según alguna regla. En este tipo de búsqueda se pretende usar el carácter aleatorio, para no caer en mínimos locales e incluso moverse temporalmente a otros estados con peores soluciones. Además, se pueden detectar interdependencias entre atributos que la búsqueda heurística no captaría. Normalmente, el criterio de parada es un número de iteraciones, y no se pueden tener garantías de que el subconjunto elegido sea el óptimo.

1.5.2.5. Medidas de evaluación de atributos

Las medidas para la evaluación de atributos suelen ser variadas. Las principales suelen estar basadas en distancia, consistencia, relevancia entre atributos. También se encuentran las métricas basadas en la información y la exactitud. En el aprendizaje supervisado, el principal objetivo de un clasificador es maximizar la exactitud en la predicción de nuevos ejemplos, esto hace que la exactitud sea aceptada y muy utilizada como medida de evaluación. Principalmente los algoritmos que emplean la exactitud como medida de evaluación de los atributos, son los de filtro y los envolventes.

1.5.2.6. Objetivos a optimizar

Los algoritmos de selección de atributos según Sánchez (2005), se pueden dividir en tres grupos conforme al objetivo que se desea optimizar: 1) los algoritmos que buscan un subconjunto con un tamaño especificado que optimice un criterio de evaluación dado, 2) aquellos métodos que localizan el subconjunto de menor tamaño que satisfaga una cierta restricción sobre el criterio de evaluación, y 3) los que intentan encontrar un compromiso entre el tamaño del subconjunto y el valor de su criterio de evaluación.

1.6. Conclusiones del capítulo

En este capítulo se han incluido los elementos teóricos de carácter general que deben ser considerados en el aprendizaje supervisado, en la reducción de la dimensión de los datos y en la selección de atributos. No se ha detallado ningún algoritmo de selección de atributos en específico. En este sentido se recomienda al lector las referencias Guyon y Elisseeff (2003); Sánchez (2005) donde se describen las principales aportaciones.

2. Diseño del algoritmo de selección de atributos

En este capítulo se aborda sobre los aspectos de diseño del algoritmo de selección de atributos, que en lo adelante se le denominará AHSA[1]. En este sentido se introducen elementos de la descomposición matricial *CUR* que es el método básico del algoritmo. Se presenta la estructura y se describen cada uno de los pasos del AHSA. También se incluye una descripción elemental de la implementación en lenguaje R del algoritmo AHSA.

2.1. Descomposición matricial CUR

La descomposición matricial *CUR* resulta una herramienta interesante, cuya utilización puede ser un auxiliar importante en la interpretación cabal de los resultados de un análisis exploratorio de datos Mahoney y Drineas (2009). A continuación se describe brevemente en qué consiste esta técnica. La descomposición matricial *CUR* Drineas y otros. (2006a,b,c), consiste en obtener una descomposición de una matriz en tres matrices de bajo rango de la manera siguiente:

$$A_{m\times n} \approx C_{m\times k} \cdot U_{k\times r} \cdot R_{r\times n}$$

tal que $k < n$ y $r < m$. La obtención de las matrices C (subconjunto de columnas de A), U (matriz de transición) y R (subconjunto de filas de la matriz A) es un caso especial del problema de selección de un subconjunto de columnas de una matriz ($min\{||A - CUR||\}$).

La descomposición matricial *CUR* propuesta para mejorar el análisis exploratorio de datos, consiste en construir C y R a partir de la determinación de un factor de importancia para cada columna

[1] **A**lgoritmo **H**íbrido para la **S**elección de **A**tributos

de la matriz de datos. Las columnas y filas de la matriz se seleccionan aleatoriamente según la distribución de probabilidad establecida por los factores de importancia, los que se interpretan como sensores de la influencia de cada columna en la mejor aproximación de menor rango de la matriz de datos. Para entender la definición de los factores de importancia dados por Mahoney y Drineas (2009), es necesario utilizar la expresión conocida de las columnas de una matriz en función de su descomposición en valores singulares. Si se denota por A_j a la j-ésima columna de A entonces se tiene que:

$$A_j = \sum_{p=1}^{r} \left(\sigma_p v_j^p \right) u^p \approx \sigma_1 v_j^1 u^1 + \sigma_2 v_j^2 u^2 + \ldots + \sigma_k v_j^k u^k$$

donde r es el rango de A, v_j^pes la j-ésima componente del p-ésimo vector singular derecho, u^p es el p-ésimo vector singular izquierdo y σ_p es el p-ésimo valor singular.

En otras palabras, se tiene que cada columna de la matriz A puede escribirse como una combinación lineal de los vectores singulares derechos. Luego, si se trata de aproximar la matriz A por una de rango menor k, entonces es válido aproximar A_j por una combinación lineal de los k vectores singulares derechos asociados a los mayores valores singulares (llamados usualmente mayores vectores singulares) obteniéndose:

$$A_j \approx \sigma_1 v_j^1 u^1 + \sigma_2 v_j^2 u^2 + \ldots + \sigma_k v_j^k u^k$$

Debe resaltarse el hecho de que en cada uno de los sumandos la dependencia de j sólo aparece en el término correspondiente a los vectores singulares izquierdos. Entonces resulta directo asociar el factor de importancia de la columna j-ésima, con la suma de las componentes j-ésimas de los k mayores vectores singulares izquierdos.

De esta forma el factor de importancia normalizado de la j-ésima columna se define como:

$$\pi_j = \frac{1}{K} \sum_{p=1}^{K} \left(v_j^p \right)^2 \tag{2.1.1}$$

el vector π de componentes π_j es un vector de distribución de probabilidad.

A partir del vector de probabilidad definido por la expresión (2.1.1), el algoritmo ColumnSelect Mahoney y Drineas (2009), selecciona columnas de una matriz de datos A con un parámetro de rango K y un parámetro de error ε.

2.1.1. Algoritmo ColumnSelect

El algoritmo ColumnSelect tiene las entradas siguientes:

La matriz A, la Descomposición del Valor Singular (SVD) Golub y Van Loan (1996), de la matriz A y k la cantidad de columnas a seleccionar .

1. Calcular $v^1, \ldots, v^k$ (los mayores k vectores singulares derechos de A) y los factores de importancia normalizados según la expresión (2.1.1).
2. Mantener la j-ésima columna de A con probabilidad $p_j = min\left(1, c\pi_j\right)$ para todo $j \in \{1, \ldots, n\}$ donde $c = O\left(k\frac{logk}{\varepsilon^2}\right)$
3. Regresar la matriz C que contiene las columnas seleccionadas de A

Mediante este procedimiento, la matriz C contiene c'columnas, donde $c' \leq c$ "en esperanza"

El resultado teórico más importante que avala el algoritmo establece que con probabilidad al menos del 99 %, esta elección de columnas satisface que:

$\| A - P_c A \|_f \leq \left(1 + \frac{\varepsilon}{2}\right) \| A - A_k \|_f$ donde P_c denota la matriz de proyección sobre el espacio columna generado por C y A_k es la matriz de rango k más próxima a A en norma de Frobenius. En Drineas y otros. (2008) se encuentra la demostración del resultado.

De esta forma el resultado garantiza que si A es una matriz cercana a una matriz de rango k, entonces, con alta probabilidad el subespacio generado por las columnas de A está próximo al subespacio generado por las columnas de C. La información contenida en la matriz C de la descomposición CUR es suficiente para los fines propuestos, pues permiten seleccionar las variables más importantes según el algoritmo ColumnSelect. En otras aplicaciones donde se necesita la descomposición CUR de forma explícita, el algoritmo ColumnSelect se aplica adicionalmente a la matriz traspuesta y se calcula $U = C^+AR^+$donde X^+denota la inversa generalizada de Moore-Penrose de una matriz. Para detalles del algoritmo general, ver en Drineas y otros. (2006a,b,c).

2.2. Elementos de diseño del algoritmo AHSA

Considérese una matriz de datos $Q_{n\times m}$ que tiene n observaciones (empresas) y m atributos (razones financieras), y un vector de clases $Y_{n\times 1}$ donde cada una de sus componentes $y_i \in \{-1, 1\}$ representa

la etiqueta de clase (insolvencia $y = -1$, solvencia $y = 1$) $i = 1, ..., n$ (ver figura (2.2.1)). Como ya se mencionó en el apartado (1.5.1) el problema de selección de atributos se formula como el problema de optimización siguiente:

$$\begin{aligned} &\text{maximizar} \quad E(Z(X)) \\ &\text{sujeto a:} \\ &X \subset Q \end{aligned} \tag{2.2.1}$$

Donde E es la función objetivo que está dada por la precisión de un clasificador Z que emplea los atributos X de $Q_{n\times m}$. La función E se define como la razón de clasificación correcta si las clases están balanceadas. En caso en que no estén balanceadas las clases en la muestra se define E, como la media de la razón de clasificación correcta de cada clase Ferri y otros. (2009).

		Razones				*Estado*
Empresas	x_1	x_2	$\cdots$		x_m	Y
Emp_1	$x_{1,1}$	$x_{1,2}$	$\cdots$		$x_{1,m}$	y_1
Emp_2	x_{21}	$x_{2,2}$	$\cdots$		$x_{2,m}$	y_2
			$\ddots$			
$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$
				$\ddots$		
Emp_n	$x_{n,1}$	$x_{n,2}$	$\cdots$		$x_{n,m}$	y_n

Figura 2.2.1.: Matriz de datos.

2.2.1. Estructura general del algoritmo AHSA

La idea central del AHSA es encontrar de manera iterativa una cantidad $p < n$ de atributos que tengan alta capacidad predictiva. El algoritmo comienza con un preprocesamiento del conjunto de datos. Luego para la solución del problema (2.2.1) el algoritmo se basa en un mecanismo de búsqueda local que emplea el algoritmo ColumnSelect para la generación de soluciones (subconjuntos de atributos $X_{m\times k}$). Para evaluar cada solución se emplea la precisión de la clasificación realizada con $T_{m\times q}$ componentes ortogonales calculadas mediante la regresión *PLS* o *PCA* a partir de los atributos $X_{m\times k}$. Para clasificar se emplea el *LDA* o el método *kNN*. Para verificar la clasificación se puede

emplear Validación Cruzada, *LOO* o Bootstrap. En la figura (2.2.2) se muestra un esquema de la estructura de una iteración del algoritmo AHSA.

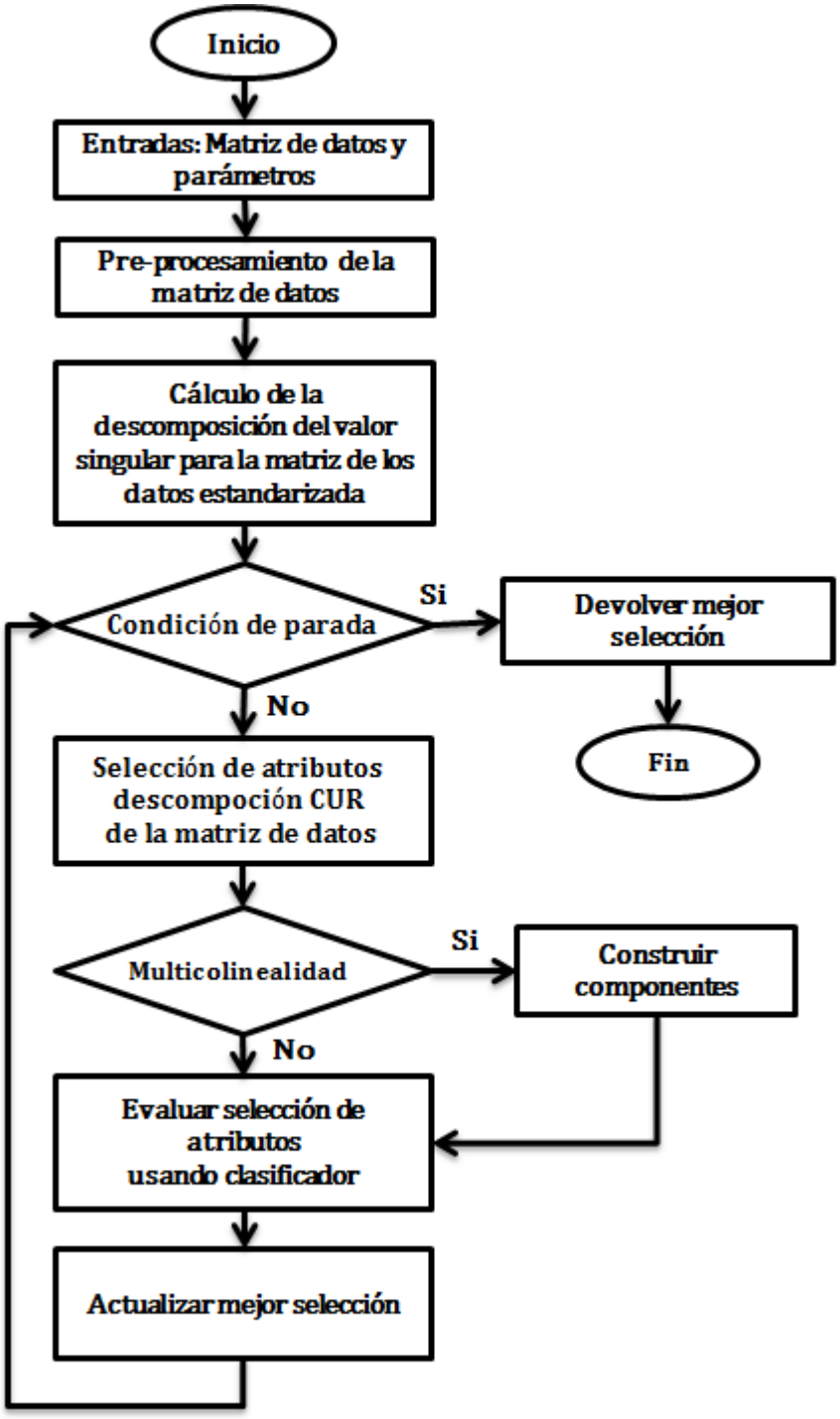


Figura 2.2.2.: Esquema del algoritmo AHSA.

A menudo las matrices de datos formadas por razones financieras presentan multicolinealidad, debido a la expresión con que se calculan. Por ejemplo, dos o más razones pueden coincidir en el denominador o el numerador. La multicolinealidad puede afectar el desempeño de algunos clasificadores, por esta razón en el algoritmo propuesto (ver figura (2.2.2)), se emplean las componentes en lugar de los atributos originales, cuando estos van a ser evaluados mediante la precisión del clasificador.

2.2.2. Caracterización del algoritmo AHSA

Al AHSA se le concede la denominación de algoritmo híbrido pues combina la idea de los algoritmos envolventes y los de filtro. El AHSA es un algoritmo envolvente que tiene un mecanismo de filtro aleatorio, para la generación de los subconjuntos de atributos a evaluar en cada iteración. Se dice que el AHSA emplea un algoritmo de filtro, en el sentido del uso que se le da a la descomposición CUR, la cual permite en cada iteración filtrar de manera aleatoria las columnas de la matriz de datos (atributos) a partir del factor de importancia (expresión (2.1.1)). Los atributos con mayor factor de importancia son más propensos a estar presentes en la selección que se lleva a cabo en cada iteración. El comportamiento aleatorio del algoritmo ColumnSelect hace que cada subconjunto de atributos a evaluar sea diferente. Por la característica de aleatoriedad en cada selección de un subconjunto de atributos, se puede caracterizar el AHSA como un algoritmo con estrategia de búsqueda aleatoria.

2.2.3. Descripción del algoritmo AHSA

En esta sección se describen en detalle cada uno de los pasos del algoritmo AHSA.

1. **Entradas:** Q matriz de datos, Y clases, k_{CUR} cantidad de atributos a seleccionar mediante el algoritmo ColumnSelect (2.1.1), k_{comp} cantidad de componentes a construir mediante *PCA* o *PLS* para ser empleadas en la clasificación ($k_{comp} \leq k_{CUR}$), CO_{met} método para construir las componentes ortogonales, *Class* método de clasificación (*LDA* y *kNN*). Como condición de parada el algoritmo usa $IterMax$ el número máximo de iteraciones a realizar, también se emplea ε la tolerancia para el valor de E .

2. **Preparación de los datos**: en la iteración $t = 0$ se comienza estandarizando la matriz de datos Q. Esta acción se lleva a cabo porque las razones financieras suelen estar en diferentes niveles, varios de los trabajos revisados estandarizan los datos antes de aplicar los métodos de clasificación Du Jardin (2009); Chen y otros. (2011); Yang y otros. (2011); Serrano y Gutiérrez (2013). La estandarización es un tratamiento que usualmente se le aplica a los datos en cualquier aplicación de técnicas de *DM*. Consiste en centrar la matriz de datos, restando las medias de las columnas a cada uno de los valores de la columna correspondiente, es decir:

$$X_j = x_{ij} - \overline{X_j},\ i = 1,...,n \text{ y } j = 1,...,m$$

Donde $\overline{X}_j$ es la media de la columna j. El segundo paso en la estandarización, es el escalado de la matriz de datos. Este se realiza una vez que la matriz Q ha sido centrada. Consiste en dividir los elementos de cada columna (centrada) por el valor de la desviación típica de dicha columna, lo que sería:

$$X_j = \frac{x_{ij}}{\sqrt{\frac{\sum_{i=1}^{n} x_{ij}^2}{n-1}}},\ i = 1,...,n\ i = 1,...,n \text{ y } \quad j = 1,...,m$$

Otro preprocesamiento que se le realiza a la matriz de datos es la detección de la multicolinealidad. Al detectar que no existe multicolinealidad, el algoritmo AHSA en lugar de usar en la clasificación las componentes $T_{m\times q}$ se emplean los atributos originales. La multicolinealidad describe la dependencia lineal entre los atributos. Es un problema que hace difícil cuantificar con precisión el efecto que cada atributo ejerce sobre la variable dependiente. El número de condición es un indicador de multicolinealidad global de los atributos de un conjunto de datos y se calcula mediante la expresión:

$$\eta = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Donde λ_{max} y λ_{min} son los autovalores máximo y mínimo de la matriz de correlaciones entre los atributos. También como regla general, si $\eta \geq 30$, entonces existe multicolinealidad Vega-Vilca y Guzmán (2011).

3. **Descomposición del Valor Singular:** calcular la *SVD* de la matriz de los datos estandarizada $[U,D,V] = svd(Q)$ Golub y Van Loan (1996). La *SVD* se utiliza como entrada en el algoritmo ColumnSelect.

4. **Condición de parada:** el algoritmo AHSA cuenta con un ciclo principal que se ejecuta mientras que no se cumpla una de las dos condiciones de parada. La condición 1 es la cantidad

máxima de iteraciones y la condición 2, es que la evaluación del mejor subconjunto de atributos seleccionado sobrepase un nivel de tolerancia ε determinado.

5. **Selección de un subconjunto de atributos:** la selección de un subconjunto de atributos en cada iteración se lleva a cabo mediante el algoritmo ColumnSelect (ver (2.1.1)). Cada subconjunto X de atributos está formado por k_{CUR} columnas (atributos) de la matriz de datos estandarizada.
6. **Construcción de componentes:** la construcción de componentes se realiza si la matriz de datos estandarizada presenta multicolinealidad. En este caso se construyen k_{comp} componentes T a partir de los atributos X, para ser usadas en el clasificador. Para calcular las componentes se puede emplear el *PCA* o *PLS*. Si los datos no tienen multicolinealidad no se calculan las componentes, usándose directamente los atributos X en el clasificador.
7. **Evaluar el subconjunto de atributos:** el subconjunto de atributos X seleccionado en cada iteración es evaluado por su capacidad predictiva. La cual es estimada usando un clasificador y un método de estimación (Validación Cruzada, *LOO* y Bootstrap). Para clasificar se emplea el *LDA* o *kNN*.
8. **Actualizar la mejor selección:** en la iteración $t = 0$ se toma el subconjunto de atributos actual como el encontrado. En las sucesivas iteraciones se actualiza la mejor selección de atributos en caso que la selección de la iteración en curso, mejore la archivada.
9. Ir al paso 4.
10. **Ranking de atributos:** en cada iteración t se archivan ϕ_t^j (frecuencia de selección del atributo t) y E_t^j (evaluación del subconjunto de atributos en que ha estado presente el atributo j en todas las iteraciones realizadas por el algoritmo). Para construir el ranking se calcula Γ_j^t según la expresión siguiente:
    $\Gamma_j^t = \frac{1}{2}\left(\frac{\phi_j^t}{IterMax} + \left(\frac{E_t^j}{\phi_t^j}\right)\right),\ j = 1,...,m,\ t = 1,...,t_{max}$
    Una vez ejecutadas todas las iteraciones del algoritmo se ordenan los atributos según el valor de Γ_j^t en forma descendente. Con este ranking se prioriza a los atributos de acuerdo a su frecuencia de aparición en el subconjunto X y por la calidad que tienen los subconjuntos que lo integran.
11. **Salidas:** el algoritmo devuelve X_{best} el mejor subconjunto de atributos encontrados y su eva-

luación $E(X_{best})$, también devuelve el valor Γ_j^t y los atributos ordenados.

A continuación se incluye el seudocódigo del algoritmo AHSA:

Algoritmo 2.1 Algoritmo AHSA:

Entradas: Q, Y, k_{CUR}, w, met_{CUR}, t_{max}, tol, $E_{best}=0$, $t=0$
Escalar y centrar la matriz Q, y η
Calcular $[U,D,V]=SVD(Q)$
while $t<t_{max}$ & $E_{best}<tol$, $t=t+1$
  $X(t)=ColumnSelect(Q,[U,D,V],k_{CUR},met_{CUR})$
  Si $\eta \geq 30$ **entonces** $T(t)=comp(X(t),[PCA,PLS])$
    Evaluar $E(t)=class(T(t),(LDA,k-NN))$
  En otro caso evaluar $E(t)=class(X(t),(LDA,k-NN))$
  Si $E_{best}>E(t)$ **entonces** $X_{best}=X(t)$, $E_{best}=E(t)$ **end**
  Archivar ϕ_j^t y E_t^j
  end (while)
Calcular Γ
Devolver $X_{best}=X(t)$, $E_{best}=F(t)$, Γ **Fin**

2.3. Elementos de la implementación del método en el lenguaje R

El AHSA se programó en R versión 3.0.2. El R, es un lenguaje y entorno de programación para el análisis estadístico y gráfico. Dispone de funciones básicas para el análisis descriptivo de datos, análisis multivariado, inferencia estadísticas, estimación y *DM*. Provee al usuario de gran variedad de técnicas estadísticas y una potente visualización gráfica. Para la implementación del AHSA se usaron funciones de diversos paquetes que están aceptados por la comunidad científica, pues están disponibles en la web oficial del proyecto R. A continuación se incluyen las principales funciones usadas para la implementación del método.

2.3.1. Preprocesamiento

En el preprocesamiento para la estandarización de la matriz de datos se utiliza la función siguiente:

```
Paquete: {base}
Función: scale(Q,center=TRUE,scale=TRUE)
```

La función *scale* está incluida en el kernel del R.

2.3.2. Descomposición del Valor Singular

Para calcular la *SVD* de la matriz de datos se emplea la función siguiente:

```
Paquete: {base}
Función: svd(Q,nu= min(n,m),nv = min(n,m))
```

La función *svd* está en el kernel del R.

2.3.3. Algoritmo ColumnSelect

Para el algoritmo ColumnSelect se emplea la función siguiente:

```
Paquete: {rCUR}
Función:  CUR(Q,c=NC,r=dim(Q)[1],k=NC,sv=svd(G),method)
```

La función *CUR* no está presente en el kernel del R, pero sí está en la web oficial del proyecto que desarrolla el lenguaje R, la función es desarrollada por Andras y otros. (2012).

2.3.4. Construcción de componentes

Para la construcción de las componentes se emplean dos funciones:

Para el *PCA* se emplea la siguiente función que está incluida en el kernel del R y su funcionamiento está basado en Mardia y otros. (1979).

```
Paquete: {stat}
Función: princomp(Q)
```

Para la construcción de componentes mediante *PLS* se emplea la función basada en Boulesteix (2004); Boulesteix y Strimmer (2007), la cual usa el algoritmo SIMPLS de de Jong (1993), este paquete no está incluido en el kernel del R, pero sí está contenido en la página oficial del R.

```
Paquete: {plsgenominc}
Función: pls.regression(X,Y,data=G, CV = TRUE)
```

2.3.5. Clasificadores

Para la clasificación se emplean las siguientes funciones:

Clasificador *LDA*:

```
Paquete: {MASS}
Función: lda(X,Y,data=G,CV = TRUE)
```

El paquete MASS está incluido en el kernel del R y su funcionamiento está basado en Venables y Ripley (2002). Esta función tiene la opción de ejecutar la Validación Cruzada.

El clasificador *kNN* es tomado de la función siguiente:

```
Paquete: {class}
Función: knn.cv(train=Q,cl,k = 1,l = 0,prob = FALSE,use.all = TRUE)
```

Para ejecutar la estimación mediante Bootstrap se emplea la función siguiente:

```
Paquete: {boot}
Función: boot(data = YX,statistic = error.rate,R,formula = Y~X)
```

El paquete boot está incluido en el kernel del R.

3. Resultados computacionales

El algoritmo AHSA tiene componentes aleatorios, por esta razón el análisis que se realiza en este capítulo se hace de manera similar a como se procede con una metaheurística al resolver un problema de optimización. Para un mejor entendimiento se incluye: la descripción de las configuraciones del AHSA, las características de los conjuntos de datos empleados, los resultados obtenidos tras varias ejecuciones del algoritmo y la comparación con otros métodos reportados en la literatura.

3.1. Experimentación

Encontrar conjuntos de datos procedentes de la actividad contable de una empresa es sumamente difícil, ya que estos en muy pocas circunstancias son de libre acceso Kainulainen y otros. (2011). Tras una ardua búsqueda resultó imposible obtener un conjunto de datos de empresas nacionales. Por tal razón, para la obtención de datos se tuvo que recurrir a la colaboración de investigadores extranjeros que desinteresadamente respondieron a nuestro llamado. Este es el caso del Dr. Wieslaw Pietruszkiewicz[1]. Otro conjunto de datos (conjunto de datos de Du Jardin) que será usado en este trabajo están públicos en http://research.ics.aalto.fi/eiml/datasets/financialratios.data.

3.1.1. Configuración de los parámetros

El AHSA tiene 6 variantes las cuales se describen a continuación:

- AHSA-LDA-PLS: esta variante emplea el *LDA* como clasificador y cuando se va a evaluar un subconjunto de k atributos estos se sustituyen por $k_{comp} \leq k$ componentes *PLS*.

[1]Facultad de Ciencia de la Computación y Tecnología de la Información. Universidad Técnica de Szczecin, Polonia.

- AHSA-LDA-PCA: esta variante emplea el *LDA* como clasificador y cuando se va a evaluar un subconjunto de k atributos estos se sustituyen por $k_{comp} \leq k$ componentes *PCA*.
- AHSA-LDA: esta variante emplea directamente el clasificador *LDA* para evaluar un subconjunto k de atributos.
- AHSA-kNN-PLS: esta variante emplea *kNN* como clasificador y cuando se va a evaluar un subconjunto de k atributos estos se sustituyen por $k_{comp} \leq k$ componentes *PLS*.
- AHSA-kNN-PCA: esta variante emplea *kNN* como clasificador y cuando se va a evaluar un subconjunto de k atributos estos se sustituyen por $k_{comp} \leq k$ componentes *PCA*.
- AHSA-kNN: esta variante emplea directamente el clasificador *kNN* para evaluar un subconjunto k de atributos.

Para evaluar el desempeño de cada variante del algoritmo, se realizan 25 ejecuciones para la siguiente configuración de los parámetros:

- Conjunto de datos de Pietruszkiewicz: cantidad de atributos $k_{cur} = 5$, cantidad de componentes $k_{comp} = 2$, número máximo de iteraciones $IterMax = 200$.
- Conjunto de datos de Du Jardin: cantidad de atributos $k_{cur} = 7$, cantidad de componentes $k_{comp} = 2$, número máximo de iteraciones $IterMax = 300$.

La cantidad de iteraciones máxima ($IterMax$) y la cantidad de atributos a seleccionar por el algoritmo ColumSelect (k_{cur}) se definen de acuerdo a la dimensión del conjunto de datos. En ambos conjuntos de datos para el clasificador *kNN* se fija el valor $k = 3$.

Las corridas se realizaron en una laptop VIT- P2402, con procesador Intel core i3 a 2.5 Ghz y con 2 GB de memoria Ram.

3.1.2. Métodos para establecer comparación

El desempeño del algoritmo AHSA es comparado con algunos trabajos que aparecen en la literatura y que utilizan los mismos conjuntos de datos. En la comparación también se emplean métodos clásicos para la selección de atributos que aparecen implementados en las librerías del lenguaje R. Los métodos empleados para comparar que están implementados en el R, son los siguientes:

- SV-PLS: consiste en seleccionar las variables más significativas mediante la función *varia-*

ble.selection del paquete del R *plsgnomics*, y luego construir el modelo discriminante con dichas variables Boulesteix (2004). El algoritmo ordena las variables de acuerdo con el valor absoluto del peso, que define el primer componente *PLS*. Este orden es equivalente a la ordenación obtenida con el estadístico F y la prueba con varianzas iguales Boulesteix (2004).

- Step-Class: consiste en aplicar el método *stepclass* del paquete de R *klaR*. Este método envolvente consiste en seleccionar variables de manera secuencial en dos direcciones (hacia adelante y hacia atrás), donde se emplea un clasificador determinado (*LDA* o *kNN*) para evaluar la selección de variables Venables y Ripley (2002).
- E-LL y ELM-SVM: son algoritmos de ensamblado de atributos propuestos en Kainulainen y otros. (2011).
- SVM-PLS: algoritmo que emplea *PLS* para la selección de variables y luego emplea *SVM* para clasificar Yang y otros. (2011).
- NN-LVQ: algoritmo de clasificación basado en redes neuronales Yang y otros. (2011).

3.2. Resultados de la experimentación

En esta sección se incluyen los resultados obtenidos en la experimentación para cada uno de los conjuntos de datos. Se describen las características generales de los dos conjuntos de datos. También se discute acerca del comportamiento del algoritmo ante las diferentes configuraciones.

3.2.1. Resultados para el conjunto de datos de Pietruszkiewicz

Este conjunto corresponde a 120 compañías de las cuales se conocen las razones financieras de dos años para cada empresa. Por tanto la matriz de datos tiene 240 filas y 30 columnas. De las 240 observaciones se tiene que 112 están en estado de quiebra y las restantes 128 no lo están. Las empresas que están en quiebra alcanzaron este estado entre el 2 y 5 año antes de la recolección de los datos. Antes de ejecutar el algoritmo se realiza un análisis de la matriz de correlación de Pearson, detectándose varios atributos que están fuertemente correlacionados. En la tabla (3.2.1) se muestran aquellos atributos con correlación superior a $0,95$.

VARIABLES		COEF	VARIABLES		COEF
X8	**X16**	1,000	**X26**	X27	0,9817
X17	**X20**	1,000	X26	**X28**	0,9774
X21	**X22**	0,9993	X9	**X29**	0,9712
X11	X12	0,9959	**X19**	X30	0,9664

Tabla 3.2.1.: Correlaciones fuertes para los datos de Pietruszkiewicz.

La información redundante resta efectividad a los métodos de clasificación, por esta razón se decide no considerar en el análisis los atributos siguientes: X11, X16, X19, X20, X22, X26, X28 y X29. Estos mismos atributos son eliminados en el trabajo de Yang y otros. (2011).

En la tabla (3.2.2) se muestra la media (Prec.) y la desviación estándar de la precisión (Std.), en las 25 corridas realizadas por las variantes del AHSA. También en esta tabla se incluye el tiempo promedio en segundos (Ts) para una ejecución del algoritmo.

En cada columna de la tabla (3.2.2) se incluyen las medidas de desempeño para cada uno de los métodos de validación de la clasificación empleados (*LOO* y Bootstrap). Como Bootstrap tiene un elevado costo computacional y se recomienda su uso cuando se tiene un número pequeño de observaciones, la columna correspondiente a Bootstrap en las tablas de los resultados; sólo tienen valores en las filas pertenecientes al AHSA-LDA-PLS y al SV-PLS. Por tanto, sólo se hicieron experimentos limitados con el fin de probar la implementación. También en la tabla (3.2.2) se incluye el valor de la precisión que poseen los demás métodos con los que se compara.

Método de validación	**LOO**			**Bootstrap**		
ALGORITMO	**Prec.**	**Std.**	**T(s).**	**Prec.**	**Std.**	**T(s).**
AHSA-LDA-PLS	0,7675	0,005	1,72	0,7642	0,003	367,8
AHSA-kNN-PLS	0,7867	0,013	1,01	-	-	-
AHSA-LDA-PCA	0,7603	0,008	1,70	-	-	-
AHSA-kNN-PCA	0,7872	0,014	0,95	-	-	-
AHSA-LDA	0,7574	0,007	1.72	-	-	-
AHSA-kNN	0,7869	0,009	0,90	-	-	-
E-LL (Kainulainen y otros. (2011))	0,7450	0,060	-	-	-	-
ELM-SVM (Kainulainen y otros. (2011))	0,7660	-	-	-	-	-
SV-PLS-LDA (Boulesteix (2004))	0,6959	-	0,02	0,7042	-	1,67
StepClass-LDA (Venables y Ripley (2002))	0,7598	0,010	6,89	-	-	-
StepClass-kNN (Venables y Ripley (2002))	0,7794	0,017	57,3	-	-	-
NN-LVQ (Yang y otros. (2011))	0,7833	-	-	-	-	-
SVM-PLS (Yang y otros. (2011))	0,7900	-	-	-	-	-

Tabla 3.2.2.: Comparación del desempeño de los algoritmos para los datos de Pietruszkiewicz.

En la tabla (3.2.2) se aprecia que al usar *LDA* como clasificador, las variantes (AHSA-LDA-PLS y AHSA-LDA-PCA) que emplean componentes para evaluar los subconjuntos de atributos seleccionados, encuentran subconjuntos de atributos con mayor capacidad predictiva, que la variante (AHSA-LDA) que no emplea componentes para evaluar la selección. Esto no sucede en las variantes (AHSA-kNN-PLS, AHSA-kNN-PCA y AHSA-kNN) donde se emplea *kNN* como clasificador, en estos casos los resultados son muy similares. Con el uso de *kNN* como clasificador en el algoritmo AHSA se obtienen subconjuntos de atributos con mejor evaluación, que cuando se emplea el *LDA* como clasificador.

En la tabla (3.2.2) todas las variantes del AHSA tienen un desempeño ligeramente superior que los algoritmos SV-PLS-LDA y E-LL. El desempeño de todas las variantes del AHSA también es comparable con los del algoritmo ELM-SVM y con las dos variantes de la selección secuencial StepClass-kNN y StepClass-LDA. El AHSA al usar *kNN* como clasificador, sólo tiene un desempeño ligeramente inferior al algoritmo SVM-PLS de Yang y otros. (2011). En cuanto al tiempo de ejecución, todas las variantes del AHSA son más veloces que los algoritmos StepClass.

Una cuestión relevante del AHSA es su baja desviación estándar. En la tabla (3.2.3) se muestra la evaluación de los mejores subconjuntos de atributos que construyen las variantes del AHSA. Los atributos que con mayor probabilidad pertenecen al mejor subconjunto que encuentran las variantes del AHSA son los siguientes: X1, X9, X10, X14, X15, X17 y X18. En Kainulainen y otros. (2011), los principales atributos que seleccionan los algoritmos ELL y ELM-SVM son los siguientes: X1, X9, X10 y X17.

	Atributos					
Algoritmo.	**1**	**2**	**3**	**4**	**5**	**Prec.**
AHSA-LDA-PLS	X8	X1	X6	X15	X9	0,7768
AHSA-kNN-PLS	X17	X2	X3	X9	X23	0,8142
AHSA-LDA-PCA	X18	X14	X13	X9	X27	0,7773
AHSA-kNN-PCA	X17	X5	X10	X21	X13	0,8214
AHSA-LDA	X8	X1	X10	X15	X9	0,7718
AHSA-kNN	X17	X5	X8	X1	X24	0,8052
StepClass-kNN	X13	X6	X21	X10	X9	0,8041

Tabla 3.2.3.: Mejor subconjunto de atributos que selecciona cada algoritmo en el conjunto de datos de Pietruszkiewicz.

En la figura (3.2.1) se muestra la capacidad predictiva del *LDA* al considerar parejas de atributos tomadas del subconjunto: X1, X9, X10 y X14. En las gráficas de la figura (3.2.1) una observación correspondiente a una empresa quebrada se representa con (-) y a una empresa sana con (+). También en la parte superior de cada gráfica de la figura (3.2.1), se incluye el error de clasificación incorrecta.

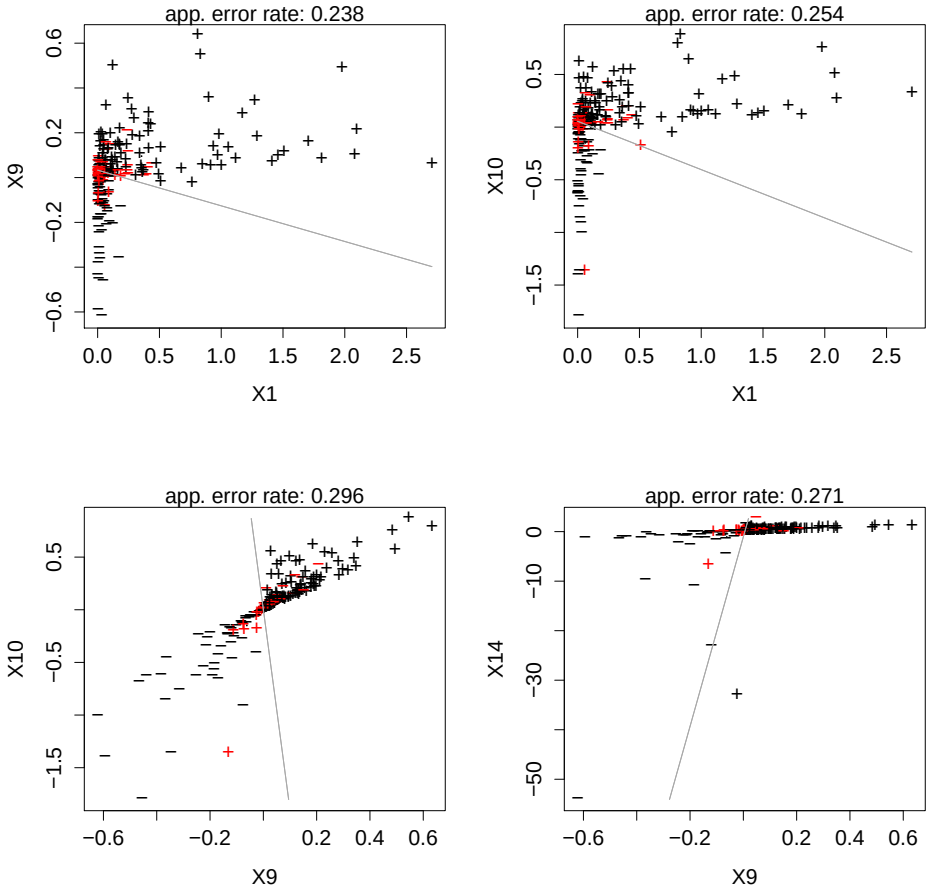


Figura 3.2.1.: Clasificación con *LDA* empleando los atributos: X1, X9, X10 y X14.

En la figura (3.2.2) se muestra la capacidad predictiva del clasificador *kNN*, al considerar parejas formadas a partir de los atributos: X5, X9, X10 y X18. Se puede observar en estas gráficas que *kNN* permite obtener mejores predicciones que el *LDA*.

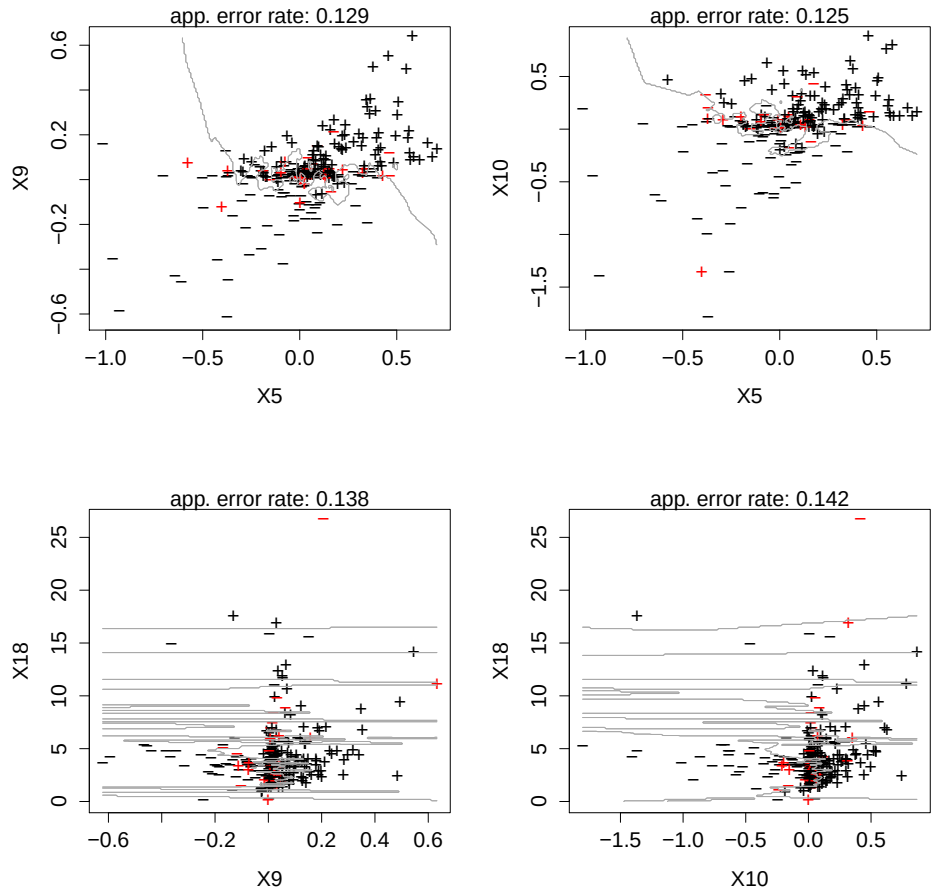


Figura 3.2.2.: Clasificación con *kNN* empleando los atributos: X5, X9, X10 y X18.

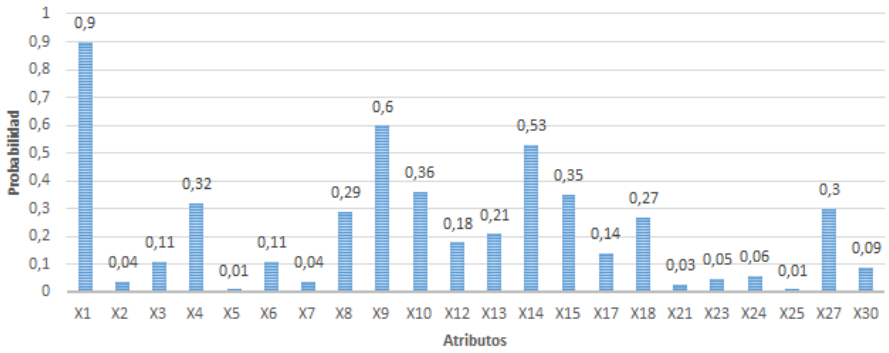
En las figuras (3.2.3) y (3.2.4) se muestra la probabilidad de selección de cada uno de los atributos del conjunto de datos de Pietruszkiewicz, por cada una de las variantes del algoritmo AHSA. Esta probabilidad se calcula según la expresión:

$$P\left(X_i \in X_{best}\right) = \frac{\sum_{t=1}^{NR} I_t}{NR}$$

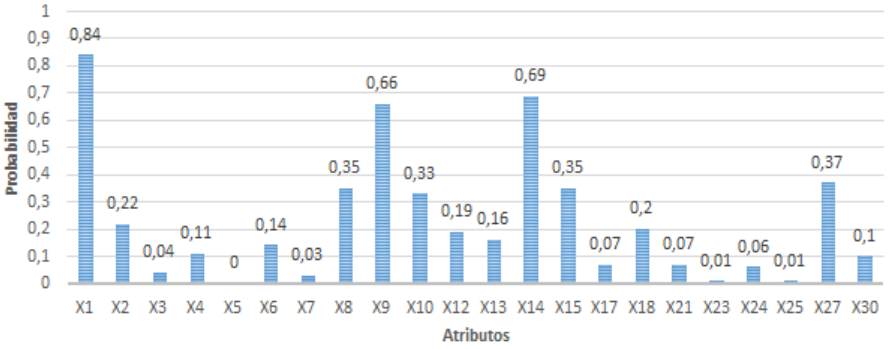
Donde $I_t = 1$ si el atributo X_i está en el mejor subconjunto seleccionado en la ejecución $t = 1,...,NR$ donde NR es el número de ejecuciones $I_t = 0$ en otro caso.

La probabilidad de selección de los atributos difiere entre las variantes del AHSA. El cambio está asociado principalmente al uso de los clasificadores *LDA* y *kNN*. Cuando se emplea el *LDA* como clasificador, los atributos que con mayor frecuencia se seleccionan son: X1, X9 y X14. Estos atribu-

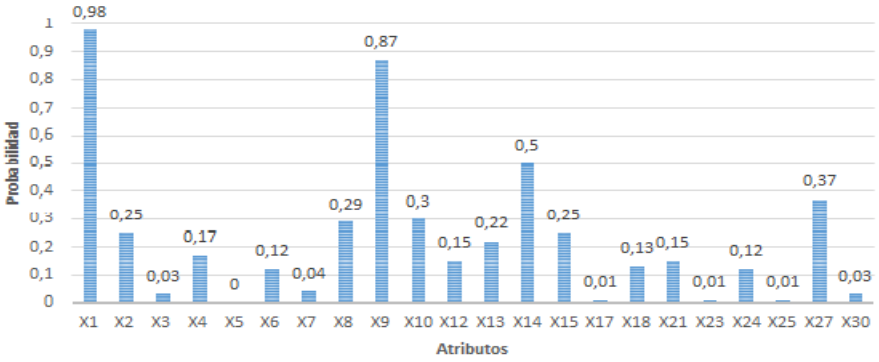
tos tienen una probabilidad de ser seleccionados superior al 0,5. También los atributos X10, X15 y X27 tienen una probabilidad considerable (>0,30).



(a) LDA-PLS



(b) LDA-PCA

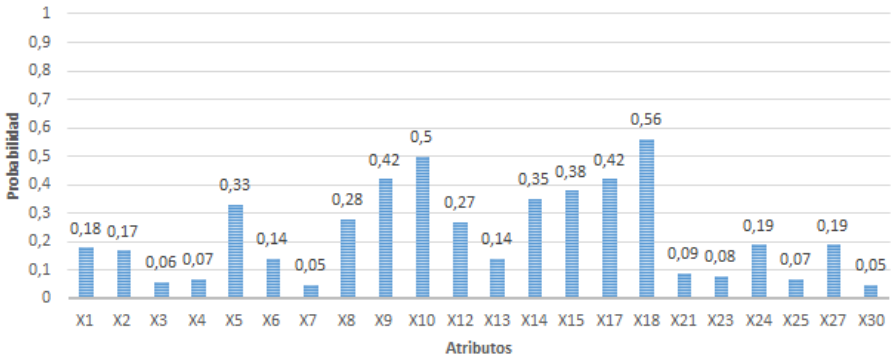


(c) LDA

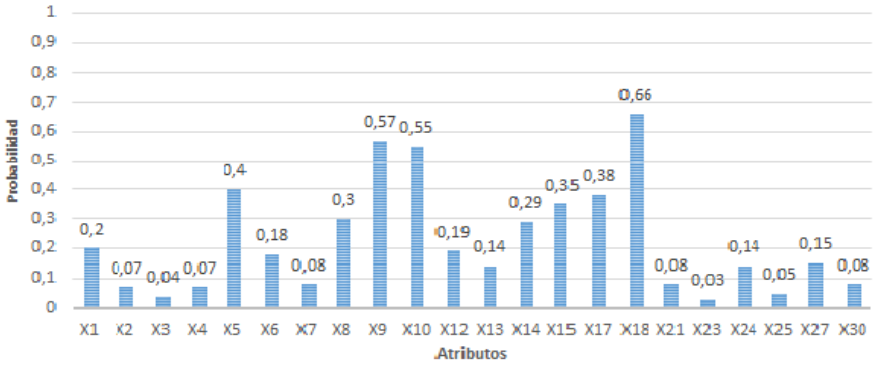
Figura 3.2.3.: Probabilidad de aparición de cada atributo en el mejor subconjunto de cardinalidad 5, que es seleccionado por cada variante del AHSA empleando el clasificador *LDA* para los datos de Pietruszkiewicz.

En el anexo (D), se incluye la capacidad de predicción de parejas de atributos que tienen probabilidad

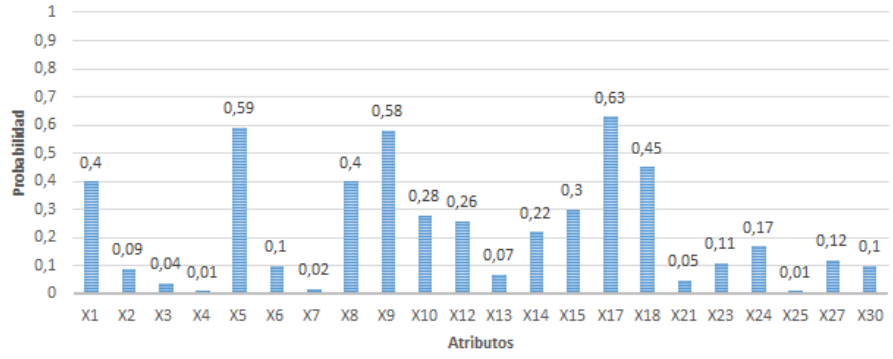
(> 0.30). En la figura (3.2.1), así como en las figuras (D.0.1 y D.0.2) del anexo (D) se puede apreciar que existen parejas de atributos que tienen una adecuada capacidad predictiva al usar el clasificador *LDA* .



(a) KNN-PLS



(b) kNN-PCA



(c) kNN

Figura 3.2.4.: Probabilidad de aparición de cada atributo en el mejor subconjunto de cardinalidad 5, que es seleccionado por cada variante del AHSA empleando el clasificador *kNN* para los datos de Pietruszkiewicz.

Al usar *kNN* como clasificador y emplear los métodos de construcción de componentes, los atributos que con mayor frecuencia se seleccionan son los siguientes: X9, X10 y X18 (ver figuras (3.2.4a) y (3.2.4b)). La probabilidad de selección de estos atributos oscila entre 0,5 y 0,66. Existen otros atributos como X5, X14, X15 y X17 que tienen una probabilidad de selección considerable que se encuentra entre 0,28 y 0,42. Al no usar componentes (figura (3.2.4c)) se mantienen con una probabilidad de selección alta los atributos X5, X9 y X18 decayendo la probabilidad del atributo X10 e incrementándose X1. Estos atributos (X1, X5, X9, X10, X14, X15, X17 y X18), al formar parejas entre sí, tienen una adecuada capacidad predictiva al usar el clasificador *kNN* (ver la figura (3.2.2) y el anexo (E)).

3.2.2. Resultados para el conjunto de datos de Du Jardin

El conjunto de datos se obtuvo de la tesis de doctorado de Fhilippe Du Jardin. Los datos proceden de 500 empresas del mismo sector con similar estructura jurídica y cantidad de activos, de las cuales se tienen registradas 41 razones financieras. En estos datos, la proporción de empresas quebradas y empresas no quebradas es de $50:50$. Antes de ejecutar el algoritmo se realiza un análisis de la matriz de correlación de Pearson, para detectar la información redundante. Solamente se detectan dos pares de variables con la correlación por encima de $0,95$ (SF1 y SF7 con $0,9869$, RE5 y RE6 con $0,9602$). Se decide no considerar en el análisis, las variables SF7 y RE6.

En la tabla (3.2.4) se muestra la media (Prec.) y la desviación estándar de la precisión (Std.), en las 25 corridas realizadas por las variantes del AHSA en el conjunto de datos de Du Jardin. También en esta tabla se incluye el tiempo promedio en segundos (Ts) para una ejecución del algoritmo. Para este conjunto de datos, los resultados son ligeramente inferiores a los algoritmos que son empleados para la comparación (E-LL, ELM-SVM, StepClass-LDA y StepClass-kNN).

También para este conjunto de datos, se observa que las variantes que emplean *PLS* para construir las componentes, tienen un desempeño ligeramente superior a las variantes que emplean *PCA*. En este caso el AHSA-LDA y AHSA-kNN que no usan componentes, tienen mejor desempeño que el AHSA-LDA-PCA.

De forma general las variantes que usan *kNN* como clasificador tienen un desempeño superior, a las que utilizan el *LDA*. Los tiempos de ejecución de las variantes de AHSA a pesar de la dimensión

de este conjunto de datos, siguen siendo mejores que los tiempos de ejecución de los algoritmos StepClass.

Método de validación	**LOO**			**Bootstrap**		
ALGORITMO	**Prec.**	**Std.**	**T(s)**	**Prec.**	**Std.**	**T(s)**
AHSA-LDA-PLS	0,9202	0,006	2,95	0,9194	0,007	691,57
AHSA-kNN-PLS	0,9210	0,006	1,86	-	-	
AHSA-LDA-PCA	0,9084	0,010	2,93	-	-	
AHSA-kNN-PCA	0,9112	0,009	1,83	-	-	
AHSA-LDA	0,9111	0,005	3,57	-	-	
AHSA-kNN	0,9210	0,006	1,98	-	-	
E-LL (Kainulainen y otros. (2011))	0,9360	0,040	-	-	-	
ELM-SVM (Kainulainen y otros. (2011))	0,9343	-		-	-	
SV-PLS (Boulesteix (2004))	0,9100	-	0,04	0,9100	-	3,31
StepClass-LDA (Venables y Ripley (2002))	0,9353	0,004		-	-	
StepClass-kNN (Venables y Ripley (2002))	0,9496	0,003		-	-	

Tabla 3.2.4.: Resultados de la ejecución de los algoritmos para el conjunto de datos de Du Jardin.

En la tabla(3.2.5) se muestran los mejores subconjuntos de siete atributos, que son seleccionados por cada una de las variantes del AHSA. En este trabajo los atributos SF1, SF2, EF3, RE3, EF2, RE5 y PR2 son los que tienen mayor probabilidad de estar en el mejor subconjunto (ver figuras (3.2.7) y (3.2.8)). En este sentido SF1 y SF2 están reportados en Kainulainen y otros. (2011) como los atributos que con mayor frecuencia aparecen en los modelos de predicción de quiebra al emplear el algoritmo E-LL.

	Atributos							
Algoritmo.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Prec.**
AHSA-LDA-PLS	PR2	EF4	EF3	RE3	SF2	EF1	SF1	0,9360
AHSA-LDA-PCA	LI8	RE2	SF1	RE5	SF10	LI5	SF2	0,9280
AHSA-LDA	RE5	EF8	LI5	SF1	RE2	EF3	SF2	0,9200
AHSA-kNN-PLS	EF2	SF1	LI5	RE4	SF2	EF3	PR1	0,9300
AHSA-kNN-PCA	RE2	SF11	EF8	RO5	RE3	SF1	EF2	0,9340
AHSA-kNN	SF11	EF2	LI8	RE5	RE3	LI5	SF2	0,934
StepClass-kNN	RE2	RE5	PR2	SF2	LI1	-	-	0,954

Tabla 3.2.5.: Mejor subconjunto de atributos que selecciona cada algoritmo en el conjunto de datos de Du Jardin.

En la figura (3.2.5), 4 gráficas muestran la capacidad predictiva del *LDA* empleando las parejas de atributos elegidas del subconjunto: SF1, SF2, RE3 y RE5. En las gráficas el signo (+) representa a la empresa sana y el (-) a la empresa quebrada. En la parte superior de cada gráfica aparece el error de clasificación incorrecta. En estas gráficas de la figura (3.2.5), se puede apreciar una elevada capacidad de predicción de los atributos que son seleccionados por las variantes AHSA.

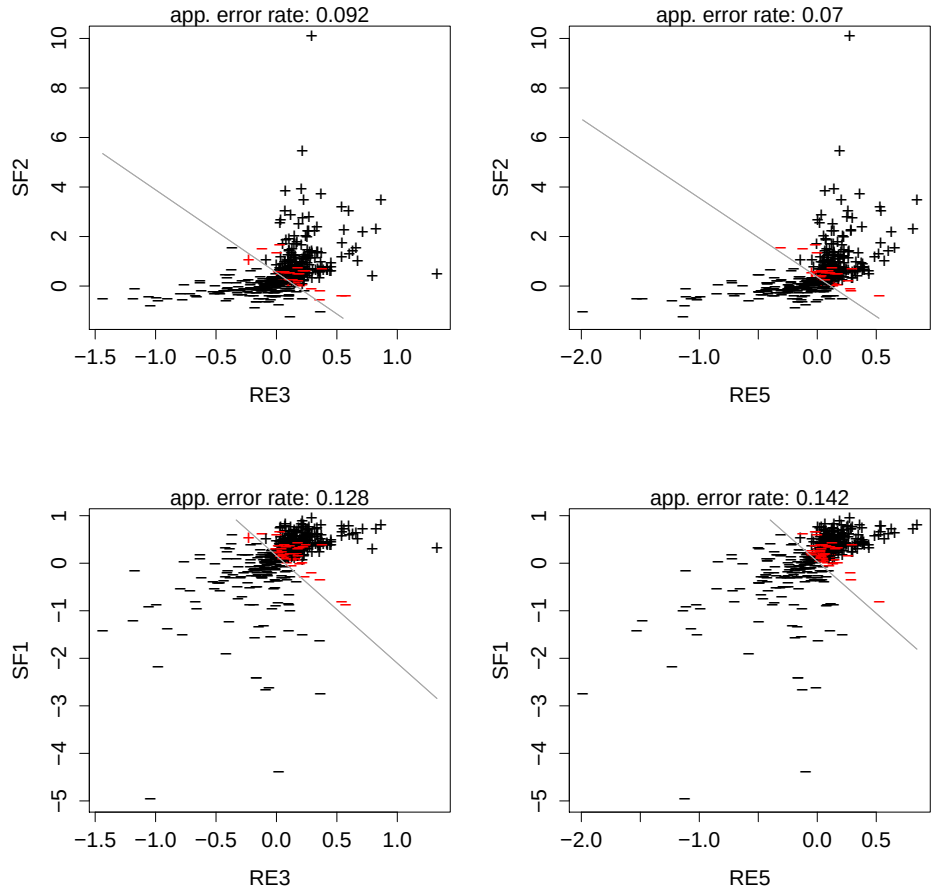


Figura 3.2.5.: Clasificación empleando *LDA* con los atributos: SF1, SF2, RE3 y RE5.

En la figura (3.2.5), se puede apreciar que existen varias parejas de atributos donde está presente la variable SF2, con las que el clasificador *LDA* tienen una precisión mayor que 0,89. En el anexo (F) también se incluyen parejas de atributos formadas con el atributo SF1, donde el *LDA* tiene una precisión entre 0,86 y 0,88.

En la figura (3.2.6) se muestra la capacidad de predicción del clasificador *kNN* al usar parejas de atributos formados a partir del subconjunto: SF1, SF2, RE3 y RE5 . Se tienen que en todos los

casos, el clasificador tiene una precisión superior a 0,95. Nuevamente se pone en evidencia que el clasificador *kNN* tiene mejor desempeño que el *LDA*, cuando son usados en el AHSA. En el anexo (G) se incluyen otras parejas de atributos, con las cuales el clasificador *kNN* obtiene un buen desempeño.

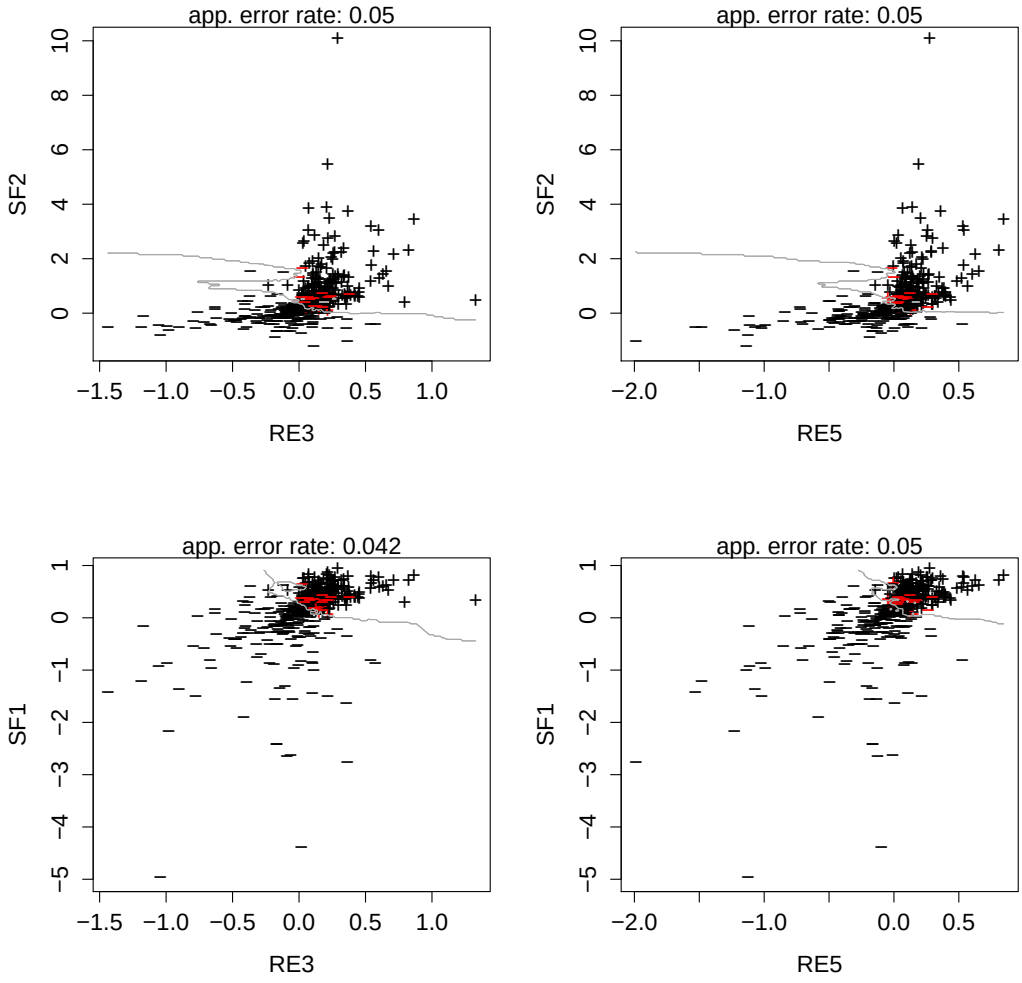
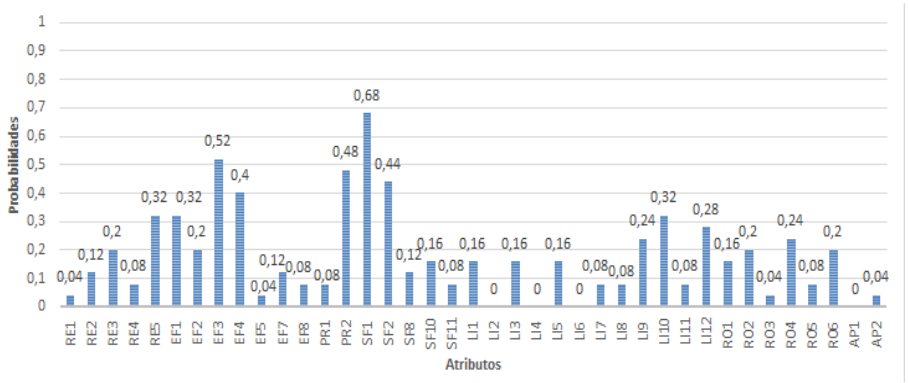
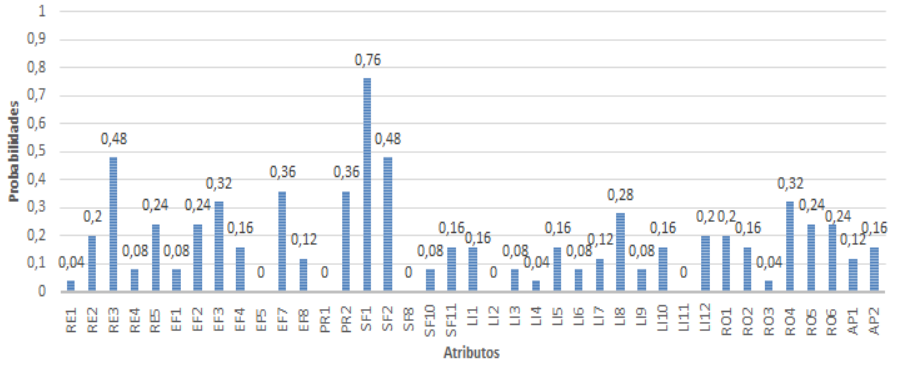


Figura 3.2.6.: Clasificación empleando *kNN* con los atributos: SF1, SF2, RE3 y RE5.

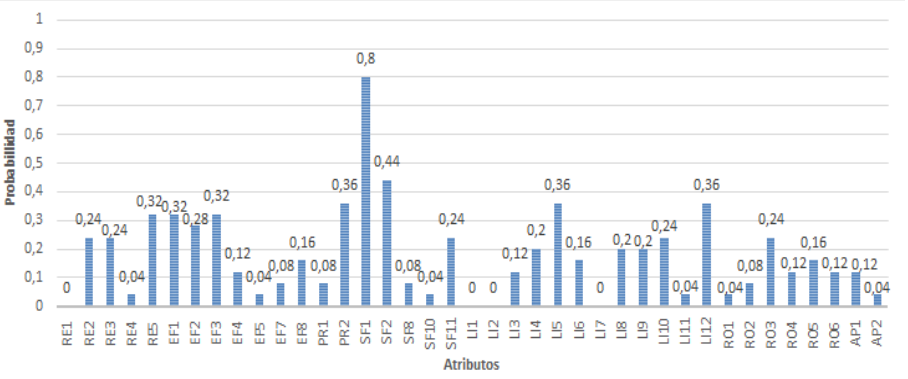
En las figuras (3.2.7) y (3.2.8) se muestra la probabilidad de aparición de cada uno de los atributos en el mejor subconjunto que seleccionan cada una de las variantes del AHSA. Se puede apreciar que los atributos con que mayor probabilidad aparecen en el mejor subconjunto, son el SF1 y el SF2 en casi todas las variantes del AHSA.



(a) LDA-PLS



(b) LDA-PCA

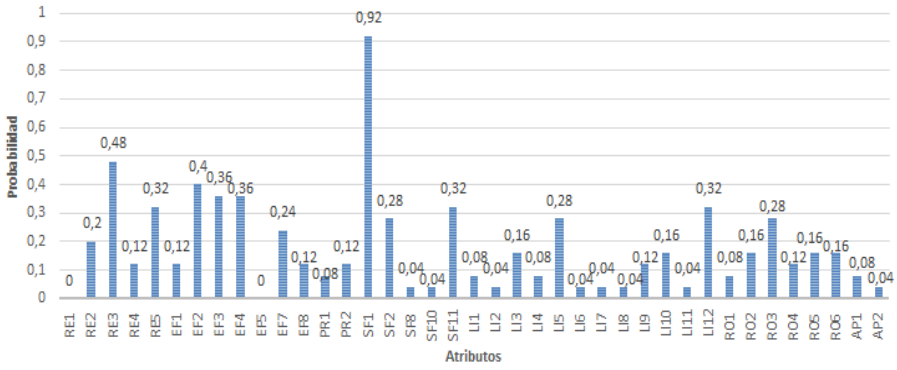


(c) LDA

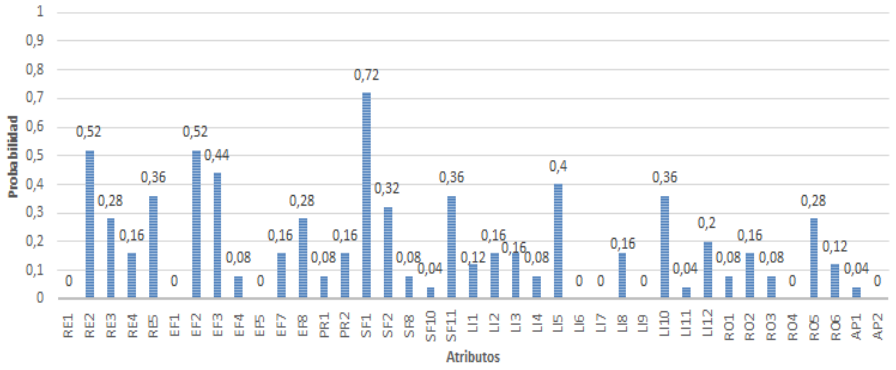
Figura 3.2.7.: Probabilidad de aparición de cada atributo en el mejor subconjunto de cardinalidad 7, que es seleccionado por cada variante del AHSA empleando el clasificador *LDA* para los datos de Du Jardin.

La probabilidad de selección de los atributos difiere entre las variantes del AHSA. Al igual que en el conjunto de datos de Pietruszkiewicz el cambio está asociado al uso de los clasificadores *LDA* y *kNN*. Cuando se emplea el *LDA* como clasificador, los atributos que con mayor frecuencia

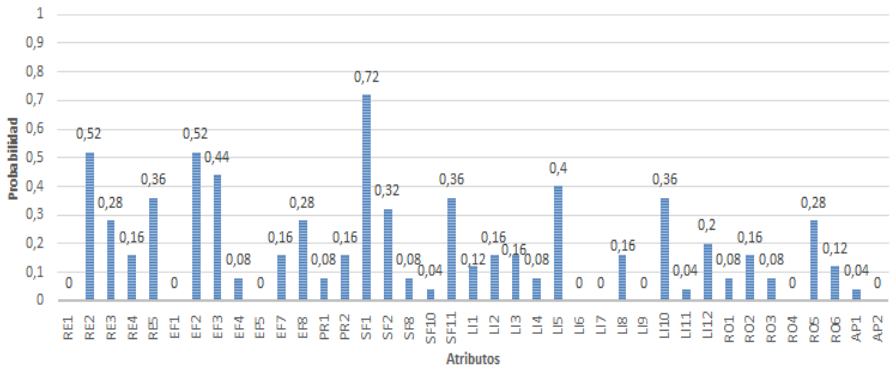
se seleccionan son: SF1, SF2 y EF3. Estos atributos tienen una probabilidad de ser seleccionados superior a 0,40. También los atributos que tienen una probabilidad considerable (> 0,30) son: PR2, RE3, RE5, EF3, EF4, EF1, LI5, LI10 y LI12 (ver Figura (3.2.7)).



(a) kNN-PLS



(b) kNN-PCA



(c) kNN

Figura 3.2.8.: Probabilidad de aparición de cada atributo en el mejor subconjunto de cardinalidad 7, que es seleccionado por cada variante del AHSA empleando el clasificador *kNN* para los datos de Du Jardin.

En este conjunto de datos, al usar *kNN* como clasificador no existe una marcada diferencia entre la probabilidad de selección de los atributos: RE3, SF1, RE2, EF2 y EF3. La probabilidad de selección de estos oscila entre 0,4 y 0,92. En el anexo (E) se muestra la capacidad predictiva de varias parejas formadas entre los mejores atributos, al usar *kNN* como clasificador.

3.3. Discusión financiera

En el trabajo de Tascón y Castaño (2012) se realiza una revisión exhaustiva de la literatura desde 1966 hasta 2009, donde se analizan las razones financieras que con mayor frecuencia se repiten en los estudios sobre fracaso empresarial y los principales factores económicos que subyacen a estas. Las razones financieras que con más frecuencia se han utilizado en los estudios son: deuda total/activo total, activo circulante/pasivo circulante, BAIT/activo total, beneficio neto/activo total, activo circulante/activo total, beneficios no distribuidos/activo total, gastos financieros/pasivo exigible y recursos generados/pasivo exigible. Debido a las variadas alternativas para formular las razones financieras que miden un mismo aspecto de la empresa (como liquidez, endeudamiento, rentabilidad, etc.) se torna difícil su interpretación. Por ello, se agrupan las razones en grupos homogéneos por su significado económico. Los grupos que propone Tascón y Castaño (2012) son los siguientes: rentabilidad, endeudamiento, equilibrio económico-financiero, estructura económica, margen y rotaciones.

3.3.1. Datos de Pietruszkiewicz

En este trabajo se decide eliminar aquellos atributos que se encuentran fuertemente correlacionados, el valor del coeficiente se puede consultar en (tabla 3.2.1). En los datos de Pietruszkiewicz fueron eliminados los atributos: X8 (ventas/activo circulante) y X16 (ventas/cuentas por cobrar), ambos permiten determinar la efectividad de la empresa, es decir, cuánto se ha generado por concepto de ventas por cada peso invertido en activos. El atributo X17 y X18 reflejan el mismo ratio (ventas/total de activos). El atributo X30 (activo corriente/ventas) y X19 (365*cuentas por cobrar/ventas) se elimina ya que X30 explica lo qué los clientes no han pagado a la empresa y el atributo X19 refleja el ciclo de cobro de los clientes de la empresa. El X21 (pasivo/total de ingreso) y X22 (pasivo co-

rriente/total de ingreso) expresan qué parte de la deuda puede ser cubierta por el total de ingresos. El atributo X26 se relaciona con los atributos X27 y X28, ambos comparten el mismo denominador (patrimonio) y el numerador varía en cuanto a la clasificación del pasivo (pasivo, pasivo a largo plazo y pasivo circulante) respectivamente. Se decide eliminar el atributo X26 y X28, en general estos expresan qué porciento de deuda es financiada con fondos propios.

Las variantes del AHSA en el conjunto de datos de Pietruszkiewicz seleccionan con mayor frecuencia los atributos: X1 (efectivo/pasivo corriente), X5 (capital de trabajo/total de activos), X9 (beneficio neto/total de activo), X10 (beneficio neto/activos corrientes), X14 (beneficio neto/patrimonio), X15 (beneficio neto/(patrimonio + pasivos a largo plazo), X17 (ventas/total de activos) y X18 (ventas/activos corrientes). En resumen el atributo X1 se ubica en el grupo de endeudamiento, X5 en el grupo de equilibrio económico-financiero y X9, X10, X14, X15 y X17 en el grupo de rentabilidad.

El subconjunto de atributos X9, X10, X14, X15 y X17 es el que provee de información sobre la rentabilidad de la empresa. En este conjunto de datos una disminución de la rentabilidad significa que la gestión operativa no marcha bien, aspecto al que se le debe prestar especial atención, para tomar acciones que reviertan esta situación lo antes posible. La pérdida de la rentabilidad aparece en la literatura como una de las fases por las que transitan las empresas quebradas Kainulainen y otros. (2011).

El atributo X5 permite evaluar el equilibrio de la solvencia y la liquidez que posee la empresa, por lo que un deterioro en este atributo supone una señal de alerta y un exceso de este puede afectar la rentabilidad. Además un deterioro en el atributo X1, afecta la capacidad de pago a corto plazo debido a una disminución de la liquidez.

3.3.2. Datos de Du Jardin

En los datos de Du Jardin, los atributos SF7 (total de deuda/total de activos) y SF1 (fondos propios/total de activos), RE6 (ingresos netos/total de activos) y RE5 (beneficios antes de intereses e impuestos/total de activos) presentan una correlación fuerte, por encima de $0,95$. Se decide eliminar los atributos SF7 y RE6 porque presentan información contenida en SF1 y RE5.

Los atributos SF1 (fondos propios/total de activos) y SF2 (deuda total/fondos propios) son los que presentan mayor probabilidad de pertenecer al mejor subconjunto encontrado por el AHSA. Estos

atributos se sitúan según Tascón y Castaño (2012) en el grupo de equilibrio económico-financiero, donde se incluyen las razones que relacionan las masas de activo con masas de financiación. Un deterioro en estos atributos, supone que la empresa presenta desequilibrio financiero, por lo que se le debe dar seguimiento a la estructura de financiamiento empleada. La relación entre la rentabilidad y la estructura financiera se destaca por la combinación entre los atributos RE3, RE5 y los atributos SF1 y SF2. Estos resultados muestran el último estadío antes de la quiebra que es el impacto de la rentabilidad en la estructura financiera, provocado por el desequilibrio económico-financiero.

4. Conclusiones

Al culminar este trabajo se puede arribar a los conclusiones siguientes:

- Con la revisión de la literatura se puede catalogar la propuesta realizada en este trabajo de investigación (AHSA) como un algoritmo híbrido, ya que en su diseño se combinaron ideas de los algoritmos de filtro y los envolventes.
- El empleo de la descomposición matricial *CUR* en el diseño del algoritmo, resultó ser acertada pues se logra confeccionar un algoritmo: eficaz en términos de la calidad de la selección y eficiente, desde el punto de vista temporal.
- La calidad del AHSA para obtener atributos con alta capacidad predictiva en la clasificación de empresas, se evalúa de forma empírica a partir de la experimentación con conjuntos de datos reportados en la literatura.
- La capacidad predictiva de los subconjuntos de atributos que selecciona el AHSA en el conjunto de datos de Pietruszkiewicz es alta, ya que es comparable con los resultados reportados en la literatura y en algunos ocasiones los sobrepasa.
- En el conjunto de datos de Du Jardin la capacidad predictiva de los subconjuntos de atributos que selecciona el AHSA, quedó ligeramente por debajo de los trabajos usados para comparar. Este comportamiento se presume que sea motivado por el clasificador y por la configuración de los parámetros realizada en la experimentación.
- El uso de las componentes *PLS* y *PCA* en el AHSA mejoran ligeramente la evaluación de los subconjuntos de atributos cuando se emplea como clasificador el *LDA*. Siendo las componentes *PLS* las que más incrementan la evaluación.
- El uso de las componentes *PLS* y *PCA* en el AHSA, cuando se emplea como clasificador *kNN*, resultó ser irrelevante en la mejora de la evaluación de los subconjuntos de atributos.

- La eficacia del algoritmo se corrobora porque el AHSA selecciona los atributos que según lo reportado en la literatura comúnmente permiten predecir la quiebra empresarial.

5. Recomendaciones

El trabajo realizado no agota todas las posibilidades de abordar el diseño de algoritmos tomando como base la descomposición matricial CUR. Otra cuestión es que en este trabajo sólo se experimentó con una sola configuración para los parámetros k_{cur} y k_{comp}. En este sentido se recomienda:

- Combinar la búsqueda secuencial con la descomposición matricial *CUR* en el diseño de algoritmos.
- Realizar experimentos para evaluar el desempeño del AHSA para otros valores de k_{cur} y k_{comp}.
- Probar el algoritmo AHSA en otros conjuntos de datos.
- Integrar el algoritmo desarrollado al módulo de obtención de indicadores financieros en Cedrux, con el fin de detectar indicadores relevantes.

Referencias

Alarcón, A. D. y Ulloa, E. I. (2012). El análisis de los estados financieros: Papel en la toma de decisiones gerenciales. *Revista académica de economía*, (167).

Alfaro, E. (2006). *Combinación de clasificadores mediante el método boosting. Una aplicación a la predicción del fracaso empresarial en España*. Tesis Doctoral, UCLM.

Altman, E. y Eisenheis, R. A. (1978). Financial applications of discriminant analysis a clarification. *Journal of financial and quantitative analysis*.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporat bankruptcy. *Journal of Finance*, 23(4):589–609.

Andras, B., Istvan, C., Michael W, M., y Norbert, S. (2012). rcur: an r package for cur matrix decomposition. *BMC Bioinformatics*, 13(103).

Archana, S. y Elangovan, K. (2014). Survey of classification techniques in data mining. *International Journal of Computer Science and Mobile Applications*, 2(2):65–71.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12(4).

Barker, M. y Rayens, W. (2003). Partial least squares for discrimination. *Journal of chemometrics*, 17(3):166–173.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4:71–127.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, 12(1):1–25.

Boulesteix, A.-L. (2004). Pls dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1075.

Boulesteix, A.-L. y Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44.

Braga-Neto, U. M. y Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380.

Brigham, E. y Houston, J. (2009). *Fundamentals of Financial Management*. Fundamentals of Financial Management. Cengage Learning.

Casanova, V. J. (2011). Revisión de los modelos de previsión de insolvencia empresarial. Tesis de Máster, Universitat Politécnica de Valéncia. Facultad de Administración y Dirección de Empresas.

Chen, N., Ribeiro, B., y Chen, A. (2011). Using non-negative matrix factorization for bankruptcy analysis. *INFOCOMP Journal of Computer Science*, 10(4):57–64.

Cover, T. y Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

Dai, J. J., Lieu, L., y Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical applications in genetics and molecular biology*, 5(1).

Dasarathy, B. V. (1991). Nearest neighbor norms pattern classification techniques.

de Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, 18(3):251–263.

Deakin, E. B. (1972). Discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(1):167–179.

Demestre, A., González, A., Del Toro, J., Arencibia, B., y Santos, C. (2005). *Análisis e Interpretación de Estados Financieros*. Centro de Estudios Contables Financieros y de Seguros.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *Systems, Man and Cybernetics, IEEE Transactions on*, 25(5):804–813.

Drineas, P., Kannan, R., y Mahoney, M. W. (2006a). Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157.

Drineas, P., Kannan, R., y Mahoney, M. W. (2006b). Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183.

Drineas, P., Kannan, R., y Mahoney, M. W. (2006c). Fast monte carlo algorithms for matrices

iii: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36(1):184–206.

Drineas, P., Mahoney, M. W., y Muthukrishnan, S. (2008,). Relative-error cur matrix decompositions. *SIAM Journal of Matrix Analysis Applications*, 30(1):844–881.

Du Jardin, P. (2009). Bankruptcy prediction models: How to choose the most relevant variables? Mpra paper, University Library of Munich, Germany.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *Man and Cybernetics, IEEE Transactions on Systems*, (4):325–327.

Efron, B. y Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Efron, B. y otros. (1979). Bootstrap methods another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Ferri, C., Hernández-Orallo, J., y Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

García, V., Herrera, C., y Ceja, J. (2012). El análisis de componentes principales aplicado a la revisión de los estados financieros de las empresas del índice habita de la bolsa mexicana de valores. *Observatorio de la Economía Latinoamericana*, (173).

Geladi, P. y Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17.

Golub, G. H. y Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 3ra edición.

Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 23:1157–1182.

Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. (1999). *Análisis multivariante*. Prentice Hall Madrid.

Hastie, T., Tibshirani, R., y Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer-Verlag.

Ibarra, A. (2001). *Análsis de las dificultades financieras en las empresas en una economía emergente: Las bases de datos y las variables independientes en el sector hotelero de la bolsa Mexicana de valores*. Tesis Doctoral, Universidad Autónoma de Barcelona.

Ibarra, A. (2009). *Desarrollo del Análisis Factorial Multivariable Aplicado al Análisis Financiero Actual*. Colombia.

Johnson, D. E. y Castellanos, J. P. (2000). *Métodos multivariados aplicados al análisis de datos*. Thomson.

Kainulainen, L., Miche, Yoan andl Eirola, E., , Yu, Q., Frénay, B., Séverin, E., y Lendasse, A. (2011). Ensembles of local linear models for bankruptcy analysis and prediction. *Case Studies in Business Industry and Government Statistics*, 4(2).

Kim, H. S. y Sohn, S. Y. (2010). Support vector machines for default prediction of smes based on technology credit. *European Journal of Operational Research*, 201(3):838–846.

Kleinberg, J. M. (1997). Two algorithms for nearest-neighbor search in high dimensions. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pág. 599–608. ACM.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95 Proceedings of the 14th international joint conference on Artificial*, pág. 1137–1145.

Kohavi, R. y John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273 – 324. Relevance.

Langley, P. (1994). Selection of relevant features in machine learning. In *In Procs. Of the AAAI Fall Symposium on Relevance*, pág. 140–144. Defense Technical Information Center.

Liu, H. y Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502.

Lo, A. W. (1986). Logit versus discriminant analysis: A specification test and application to corporate bankruptcies. *Journal of Econometrics*, 31(2):151 – 178.

Mahoney, M. y Drineas, P. (2009). Cur matrix decompositions for improved data analysis. 106(3):697–702.

Mardia, K. V., Kent, J. T., y Bibby, J. M. (1979). *Multivariate Analysis*. London Academic Press.

Nava, M. A. (2009). Análisis financiero: una herramienta clave para una gestión financiera eficiente. *Revista Venezolana de Gerencia*, 14(48):606–628.

Nguyen, D. V. y Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

Ohlson, J. A. (1980). Finanacial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pág. 109–131.

PCC (2011). *Lineamientos de la política económica y social del partido y la revolución*. La habana.

Premachandra, I., Bhabra, G. S., y Sueyoshi, T. (2009). Dea as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research*, 193(2):412 – 424.

Ribeiro, B., Silva, C., Vieira, A., y Neves (2009). Extracting discriminative features using non-negative matrix factorization in financial distress data. In Kolehmainen, M., Toivanen, P., y Beliczynski, B., editors, *Adaptive and Natural Computing Algorithms*, pág. 537–547. Springer Berlin Heidelberg.

Saeys, Y., Inza, I., y Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

Sanchis, A. (1999). *Una aplicación del análisis discriminante a la previsión de la insolvencia en las empresas españolas de seguros de no vida*. Tesis Doctoral, Universidad Complutense de Madrid, Madrid.

Serrano, C. y Gutiérrez, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems*, 54(3):1245–1255.

Siedlecki, W. y Sklansky, J. (1988). On automatic feature selection. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 2(1):197–220.

Sánchez, R. R. (2005). *Selección de atributos mediante proyecciones*. Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla.

Tascón, M. T. y Castaño, F. J. (2010). Predicción del fracaso empresarial: Una revisión. Reporte técnico, Universidad de León.

Tascón, M. T. y Castaño, F. J. (2012). Variables y modelos para la identificación y predicción del fracaso empresarial: revisión de la investigación empírica reciente. *Revista de Contabilidad*, 15(1):7–58.

Thurau, C., Kersting, K., y Bauckhage, C. (2012). *Deterministic CUR for Improved Large-Scale Data Analysis: An Empirical Study*, chapter 58, pág. 684–695.

Trygg, J. y Wold, S. (2002). Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16(3):119–128.

Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22(2):120–127.

van der Maaten, L. J., Postma, E. O., y van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71.

Vega-Vilca, J. C. (2004). *Generalizaciones de mínimos cuadrados parciales con aplicación en clasificación supervisada*. Tesis Doctoral, Universidad de Puerto Rico.

Vega-Vilca, J. C. y Guzmán, J. (2011). Regresión pls y pca como solución al problema de multicolinealidad en regresión múltiple. *Revista de Matemática: Teoría y Aplicaciones*, 18(1):9–20.

Venables, W. N. y Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer., fourth edition. edición.

Weston, J. y Brigham, E. (1994). *Fundamentos de administración financiera*. McGraw-Hill.

Wettschereck, D. y Dietterich, T. G. (1995). An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning*, 19(1):5–27.

Wold, H. (1975). Soft modeling by latent variables, the nonlinear iterative partial least square approach. *Perspectives in Probability and Statistics*, (Papers in Honour of M.S. Bartlett).

Wold, S., Sjöström, M., y Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Xiomara, V. (2005). *Nuevo enfoque para el análisis económico financiero en entidades cooperativas productoras de caña*. Tesis Doctoral, Universidad de Oriente, Santiago de Cuba.

Yang, Z., You, W., y Ji, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems with Applications*, 38(7):8336 – 8342.

Yeung, K. Y. y Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774.

Yu, L. y Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224.

A. Algoritmo SIMPLS

El algoritmo SIMPLS propuesto en de Jong (1993), es el que con mayor frecuencia se emplea en las implementaciones de los mínimos cuadrados parciales.

Algoritmo A.1 Algoritmo SIMPLS de Jong (1993).

1: $S = X_0' Y_0$
2: Para $i = 1$ hasta n
3: Si $i =$ 1, calcular la descomposición del valor singular $[u,s,v] = svd(S)$
4: Si $i > 1$, $[u,s,v] = svd\left(S - P_{i-1}\left(P_{i-1}' P_{i-1}\right)^{-1} P_{i-1} S\right)$
5: $r_i = u(:,1)$ primer valor singular izquierdo
6: Calcular componente PLS, $t_i = X_0 r_i$
7: $p_i = X_0' t_i / t_i' t_i$
8: Fin
9: Devolver $B_{PLS} = R_k\left(T_k' T_k\right)^{-1} T_k Y_0$

B. Gráfico de la matriz de correlación de Pearson de los datos de Pietruszkiewicz

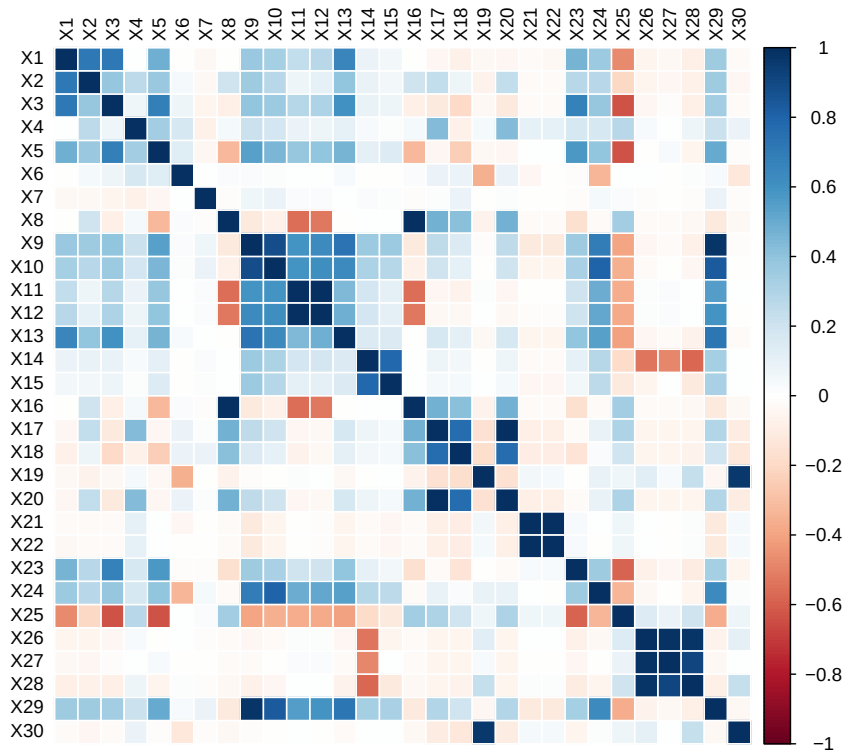


Figura B.0.1.: Matriz de correlación datos de Pietruszkiewicz.

La gráfica fue generada mediante la función:

Paquete : `ccorrplot-package {corrplot}`

Función:`corrplot(cor(Data[,1:22]),method=c("color"),order = c("hclust"), hclust.method = c("complete"),rect.col = "black")`

C. Gráfico de matriz de correlación de Pearson de los datos de Du Jardin

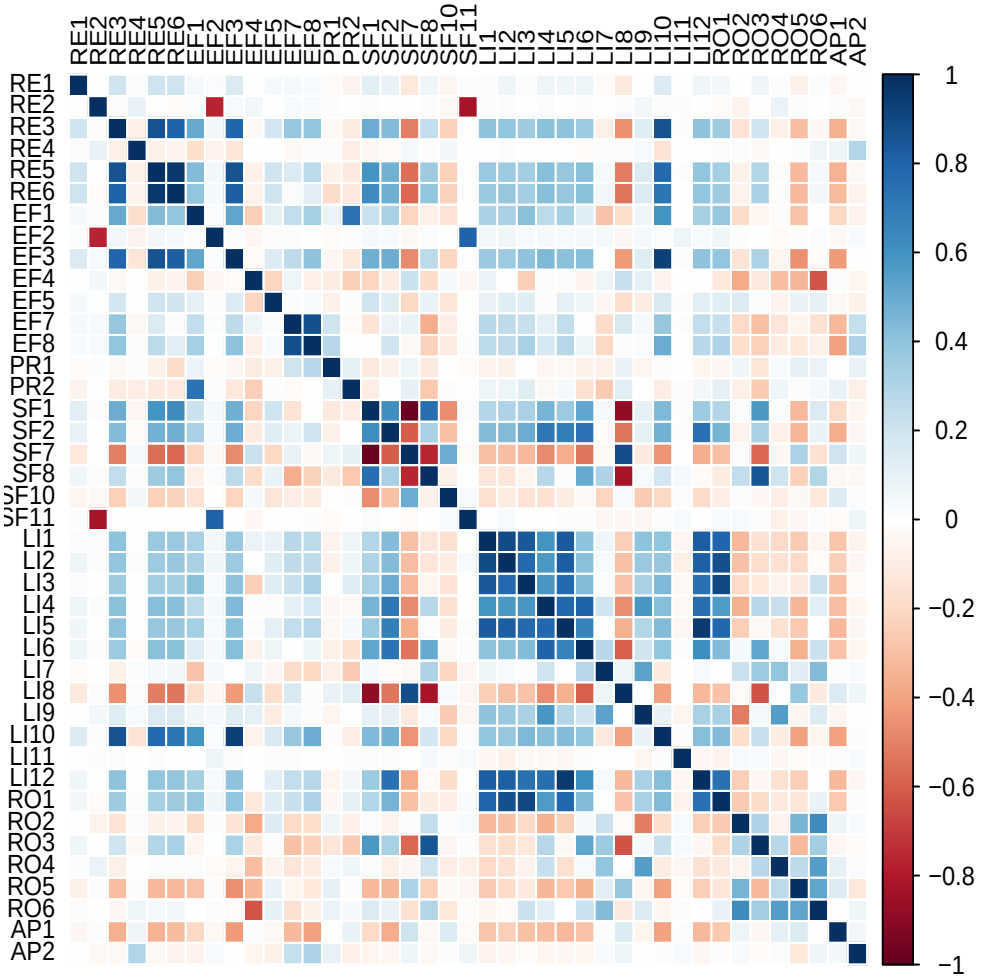


Figura C.0.1.: Matriz de correlación datos de Du Jardin.

D. LDA datos de Pietruszkiewicz

Las gráficas de clasificación fueron generadas con las siguientes funciones del lenguaje R:

Paquete: klaR

Función: drawparti(grouping=as.factor(Y),x=X[,id],y=X[,id],method="lda or k-NN",prec=200,xlab=VAR[id],ylab=VAR[id], col.correct="black",col.wrong="red",col.mean="black",col.contour="darkgray",gs=as.character(YY),pch.mean=0.1, cex.mean=0.1,print.err=1,legend.err=FALSE, legend.bg="white",imageplot =FALSE)

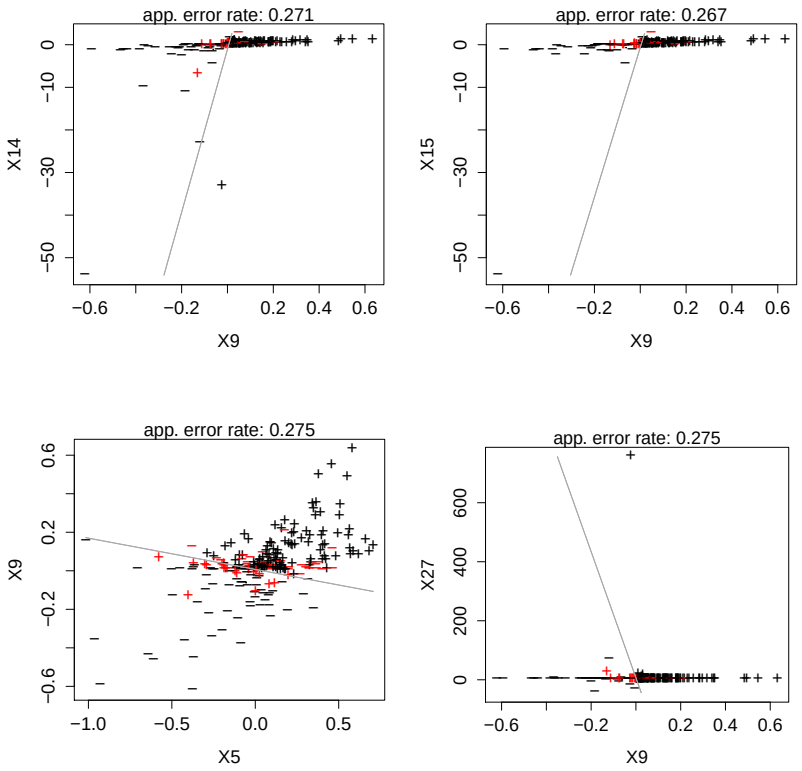


Figura D.0.1.: Clasificación empleando *LDA* con parejas de atributos tomadas del subconjunto: X5, X9, X14, X15 y X27.

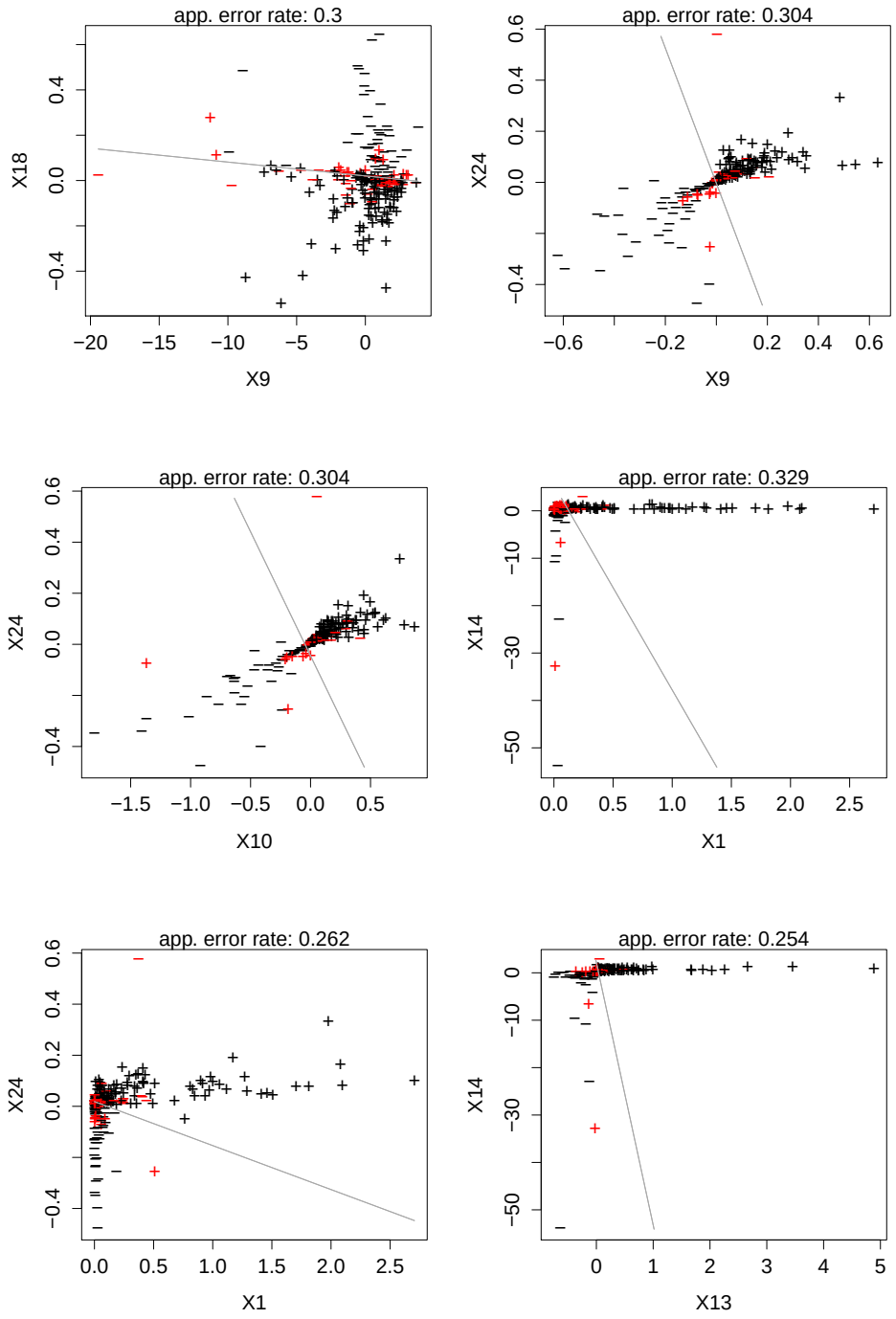


Figura D.0.2.: Clasificación empleando *LDA* con los atributos X1, X9, X10, X13, X14, X18 y X24.

E. kNN datos de Pietruszkiewicz

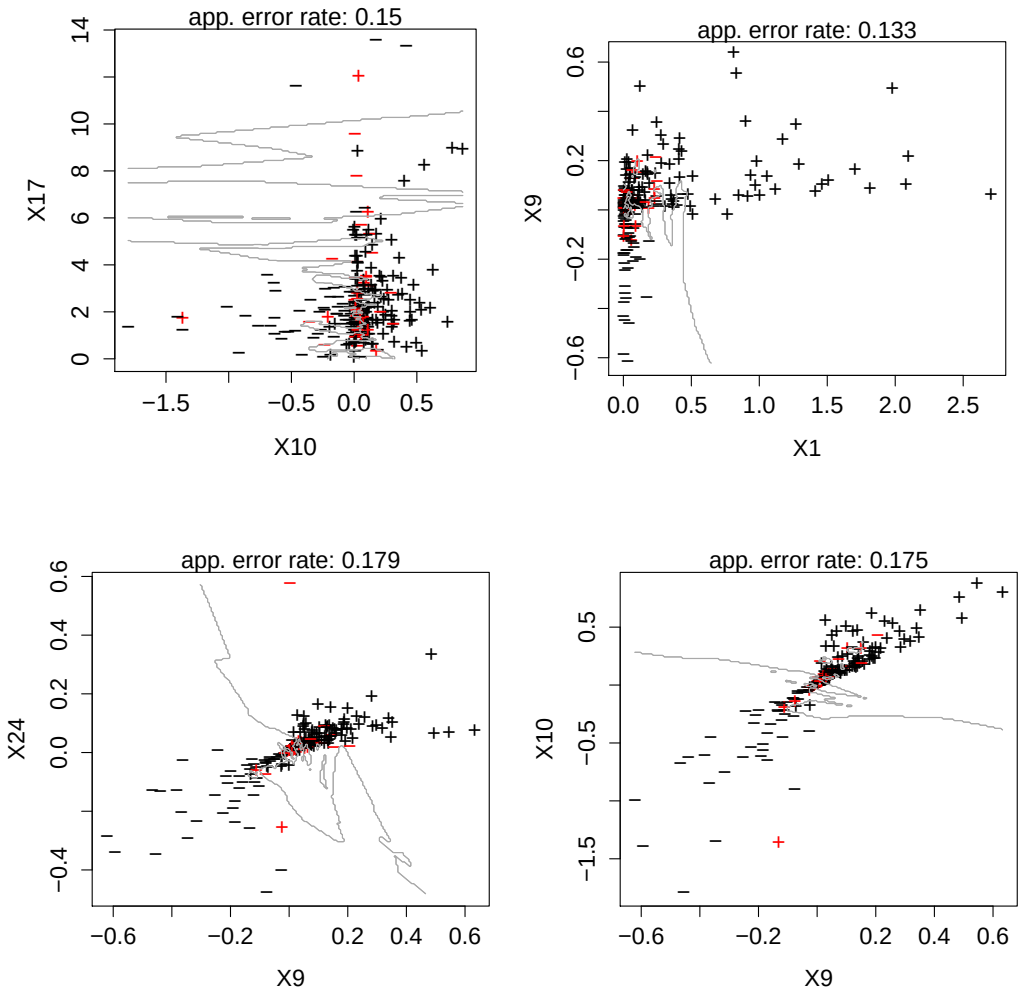


Figura E.0.1.: Clasificación empleando *kNN* con parejas de atributos tomadas del subconjunto: X1, X9, X10, X17 y X24.

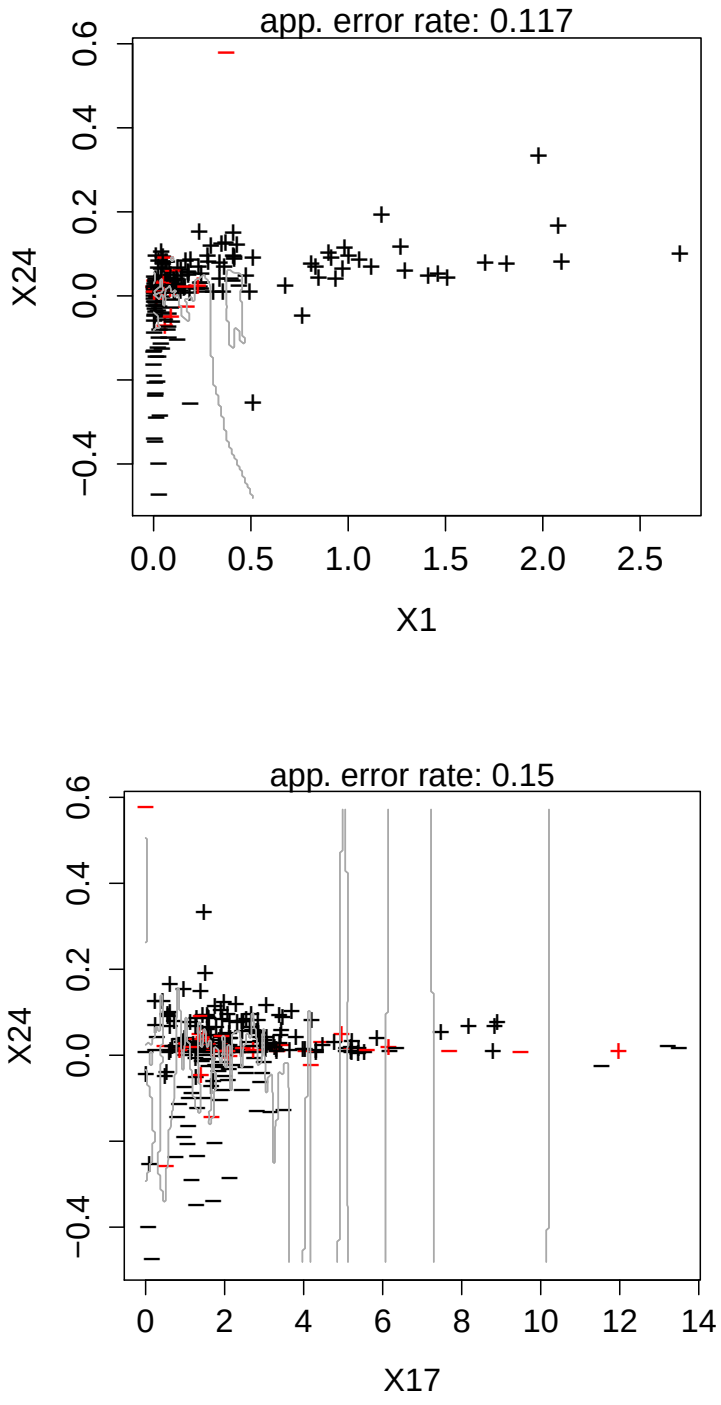


Figura E.0.2.: Clasificación empleando *kNN* con parejas de atributos tomadas del subconjunto: X1, X17 y X24.

F. LDA datos de Du Jardin

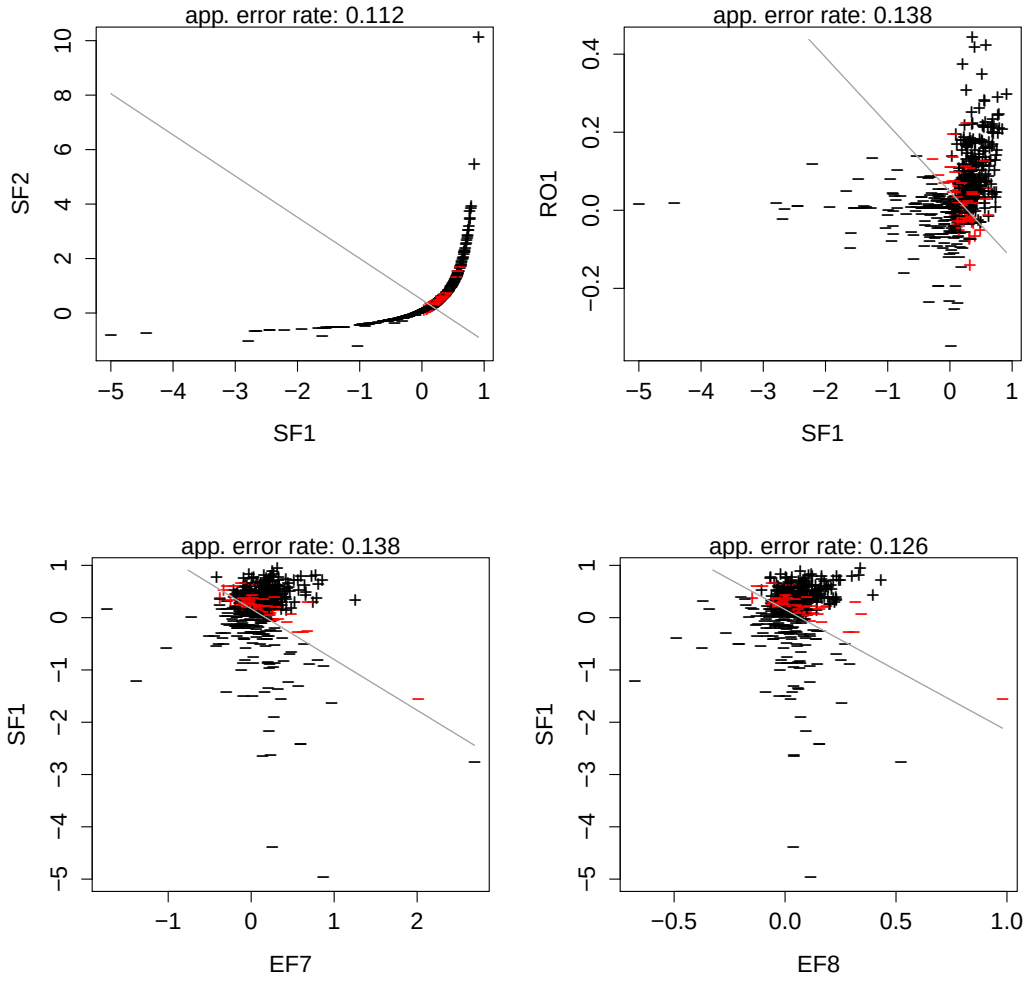


Figura F.0.1.: Clasificación empleando *LDA* con parejas de atributos tomadas del subconjunto: SF1, SF2, RO1, EF7 y EF8 .

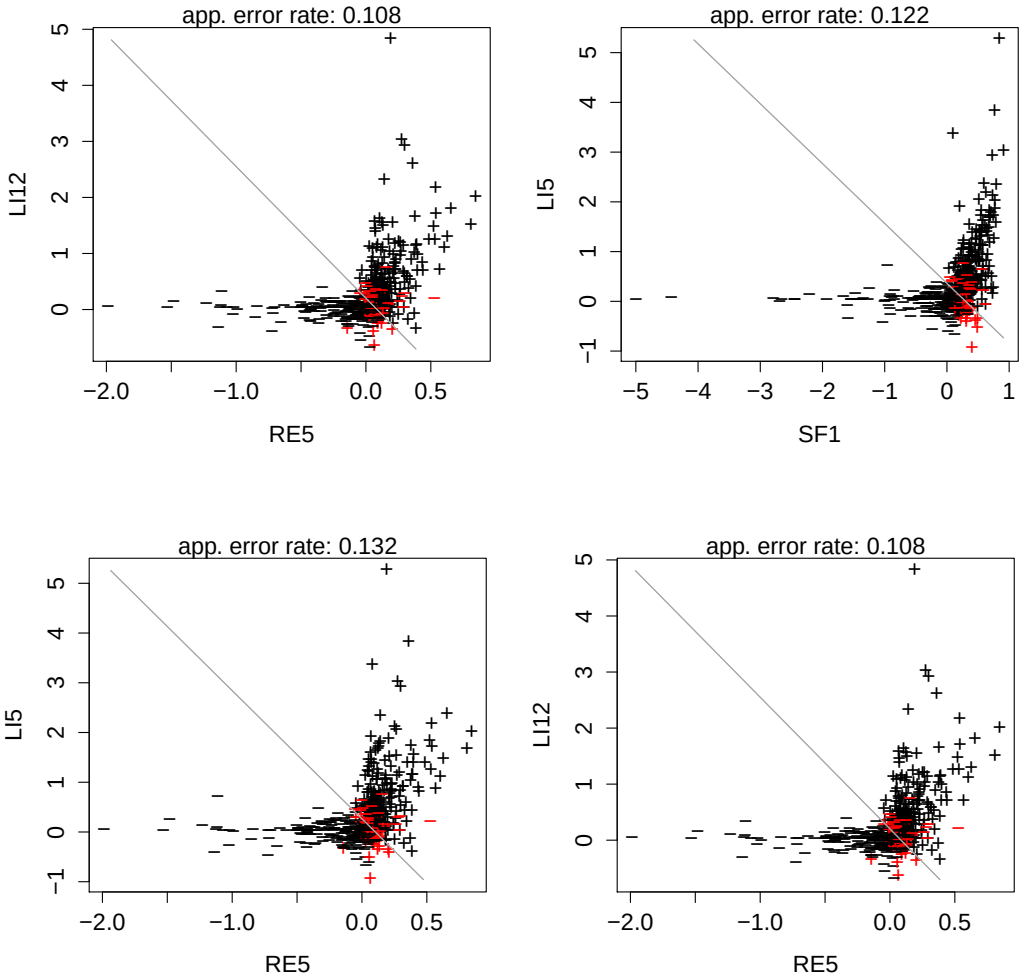


Figura F.0.2.: Clasificación empleando *LDA* con parejas de atributos tomadas del subconjunto: SF1, RE5, LI5 y LI12.

G. kNN datos de Du Jardin

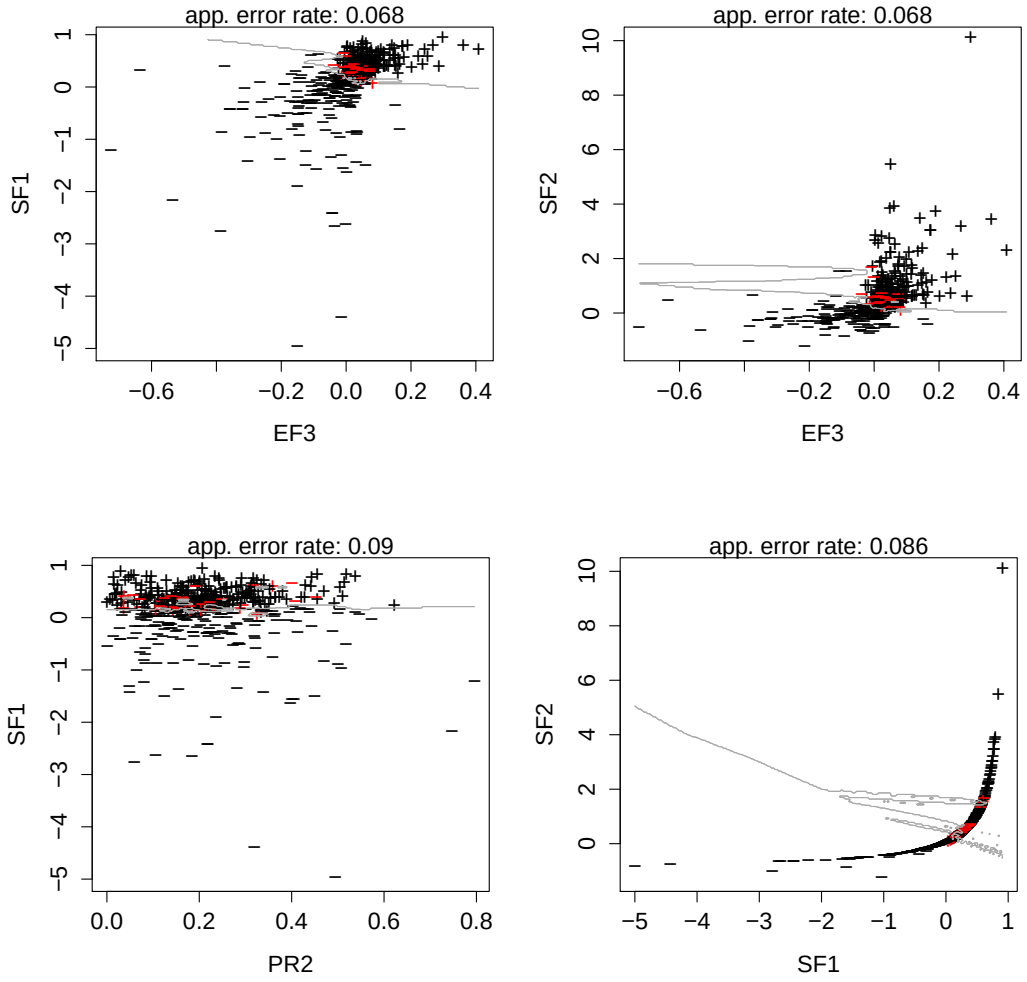


Figura G.0.1.: Clasificación empleando *kNN* con parejas de atributos tomadas del subconjunto: SF1, SF2, PR2 y EF3 .