PETICIÓN DE TEMA DE TRABAJO FIN DE

GRADO EN INGENIERÍA INFORMÁTICA

Escuela Politécnica Superior (Universidad de Córdoba)

**Paquete de selección de características en R**

Petición de tema de Trabajo Fin de Grado de:

Adán Muñoz Rodríguez

Directores:

Antonio Araúzo Azofra

# 1.- Introducción

El proyecto a realizar se centra en el uso del aprendizaje automático, una rama de la Inteligencia Artificial, que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento, es decir, un método que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares. El campo de actuación del Aprendizaje Automático se solapa con la estadística, pues ambas disciplinas se basan en el análisis de datos.

En definitiva, una de los objetivos del Aprendizaje Automático es intentar extraer conocimiento sobre algunas propiedades no observadas de un objeto basándose en las propiedades que sí han sido observadas de ese mismo objeto (o incluso de propiedades observadas en otros objetos similares).

En nuestro caso, vamos a estudiar la eficiencia de una serie de algoritmos para la selección de características de una serie de objetos dentro de una base de datos. La mayoría de las bases de datos, trabajan con volúmenes de datos muy grandes, por lo que es conveniente eliminar aquellas características menos importantes de los objetos, debido a que no aportan información relevante en el problema estudiado, o por qué son redundantes.

La selección de características es un término usado habitualmente en la minería de datos para describir las herramientas y las técnicas disponibles para reducir las entradas a un tamaño apropiado para su procesamiento y análisis. La selección de características, no solo implica la reducción del número de características a considerar un modelo, sino también la elección de las características que aportan más información (aquellas más importantes). Por lo que, los analistas o las herramientas de modelado, deberán seleccionar o descartar activamente aquellas características en función de su utilidad para el análisis.

Hoy en día hay muchas propuestas de algoritmos para la selección de características, implementadas en diversos lenguajes de programación, sin embargo, no todas estas propuestas se encuentran al alcance de cualquier usuario o analista estadístico que quiera usarlas. Podemos encontrar algoritmos escritos en Python, C, C++ o R, pero no hay disponible un repositorio de algoritmos cuyo objetivo principal sea la selección de características. Por lo que, para la comunidad de usuarios que pueden necesitar hacer uso de la selección de características, les sería de ayuda un paquete de algoritmos accesible mediante un lenguaje de programación conocido para ellos, como puede ser el lenguaje de programación R.

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Es probablemente el lenguaje más utilizado en investigación por la comunidad estadística, en campos como la bioinformática, biomedicina y las matemáticas financieras. Además, nos brinda la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráficos.

R proporciona bastantes herramientas estadísticas, como pueden ser los modelos lineales y no lineales, los tests estadísticos, los algoritmos de clasificación y agrupamiento... Incluso, para aquellos algoritmos computacionalmente más complejos, es posible desarrollar bibliotecas en C o C++ que serán cargadas dinámicamente, y utilizar el paradigma de la orientación a objetos.

Además, R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python. Otra de las características de R es su capacidad gráfica, que permite generar gráficos con alta calidad. R posee su propio formato para la documentación basado en LaTeX.

# La motivación principal de este proyecto, es facilitar el uso de losalgoritmos de selección de características más adecuados en cada caso, eligiendo los métodos de búsqueda y las medidas más interesantes tras una investigación previa. Para ello se realizará la elaboración de un paquete completo y que sea fácilmente accesible para cualquier usuario del lenguaje de programación R. Gracias a esto una vez sea publicado y aceptado en CRAN conforme a sus reglas y estándares, mucha gente podrá utilizar este paquete, de gran utilidad en el ámbito estadístico.

# 2.- Objetivos

El objetivo principal es la publicación de un paquete completo compuesto por una serie de algoritmos de selección de características para el lenguaje de programación R. Para ello, se deberán cumplir los siguientes objetivos:

* Selección de los métodos de selección de características más interesantes para su aplicación práctica.
* Modularización de los métodos, separando métodos de búsqueda y medidas de relevancia de las características.
* Implementación de las diferentes medidas a utilizar.
* Implementación de los métodos de búsqueda seleccionados.
* Diseñar una interfaz de usuario fácil de usar e intuitiva para alguien habituado a un entorno de R.
* Realizar una series de tests para comprobar el correcto funcionamiento de los métodos implementados.
* Formar un paquete compuesto por los diferentes algoritmos de selección de características, cumpliendo las reglas y los estándares de CRAN.
* Realizar una experimentación básica comparando los algoritmos implementados y analizar los resultados obtenidos.
* Publicación del paquete para su uso y posible mejora.

# 3.- Antecedentes

La selección de características, como dijimos anteriormente es una técnica muy utilizada en la minería de datos, para reducir las entradas de una B.D a un tamaño apropiado para su procesamiento y análisis. La selección de características, no solo implica la reducción del número de atributos que se pueden considerar al crear un modelo, sino también la elección de atributos más importantes según la utilidad de nuestro análisis. La selección de características siempre se realiza antes del entrenamiento del modelo y permite elegir automáticamente los atributos que con toda probabilidad se usarán en el mismo.

La capacidad de aplicar la selección de características es esencial para un análisis eficiente, pues los conjuntos de datos suelen contener mucha más información de la necesaria para la generación del modelo. Debemos eliminar los atributos ruidosos o redundantes, pues dificulta la detección de patrones significativos a partir de los datos. Por lo que, la selección de características ayuda a resolver el problema de tener demasiados datos de escaso valor o muy pocos datos de mucho valor.

Al ser la selección de características una técnica muy utilizada, podemos encontrar una gran diversidad de módulos de Aprendizaje Automático que incluyan algunas funciones para hacer uso de la selección de características. Cada módulo de los que podemos encontrar en diversas fuentes en Internet, incluye una serie de métodos e interfaces diferentes, han sido publicados bajo diferentes estándares, y cada uno está programado en un lenguaje de programación diferente (R puro, rcpp, Python, matlab…).

Nuestro objetivo es realizar un estudio de los módulos más interesantes que realicen funciones de selección de características, para implementar aquellos más interesantes en el lenguaje de programación R y optimizarlos si es posible, con el fin de unificar aquellos algoritmos de selección de características en un único módulo que pueda ser publicado en CRAN.

En CRAN, podemos encontrar un buen número de paquetes centrados en el aprendizaje automático. Todas las funciones que encontremos aquí, estarán implementadas en R. Hay una gran variedad de paquetes, cada uno abarca diferentes campos del aprendizaje automático y incluso varios de ellos, nosotros vamos a estudiar los paquetes más importantes, y nos centraremos en aquellas funciones de selección de características. Como meta final, queremos que nuestro paquete sea publicado en CRAN de acuerdo a sus normas, y que esté compuesto por algoritmos de selección de características, algo muy útil, pues podemos apenas hay paquetes escritos en R que se centre únicamente en ello.

Algunos de los paquetes que tocan el campo de la selección de características (que podemos encontrar en CRAN1) y que vamos a estudiar son:

* **Boruta**: Encuentra las características relevantes de las bases de datos, comparando la importancia de los atributos originales, con la importancia alcanzable al azar. La estimación se realiza usando sus copias permutadas.
* **FeaLect**: A cada atributo se le asigna una puntuación y se realizan los modelos para cada subconjunto de datos. Los atributos con la puntuación más baja deberían eliminarse, pues llevan a un sobreentrenamiento del modelo.
* **Fresa.cad**: Contiene un conjunto de utilidades para construir modelos basados en fórmulas lineales o logísticas para aplicaciones de pronóstico. Algunas utilidades que incluyen son, ajuste de datos, validación del modelo, análisis longitudinal, visualización…
* **FSelector**: Este paquete incluye funciones para la selección de características, identificando y eliminando aquellas características que proporcionen información irrelevante y redundante.
* **iFes**: En este paquete podemos encontrar algoritmos incrementales de selección de características. Usa algoritmos de regresión logística y de selección de características acelerados por NVIDIA GPU, para medir como de eficiente es el conjunto de características seleccionadas en cada conjunto de datos.
* **mRMRe**: Este paquete contiene un conjunto de funciones para calcular matrices de información de variables continuas y categóricas. También contiene funciones para la selección de características con mRMR y técnicas de ensamblado mRMR.
* **penalizedSVM**: Proporciona la utilidad de selección de características SVM utilizando funciones de penalización.

También utilizaremos como fuente de algoritmos de selección de características, la tesis de Antonio Arauzo Azofra2. En ella, podemos encontrar hasta 97 métodos de selección de características, mediante la combinación de los diferentes métodos de búsqueda y de las diferentes medidas a utilizar.

La búsqueda completa (algoritmo Focus 2), la búsqueda secuencial (algoritmos SFS, SBS), la probabilística (algoritmos LVF y LVW) o la búsqueda con meta-heurísticas (algoritmo SA, GA), son algunos ejemplos de los métodos de búsqueda que podemos encontrar en la tesis. Como medidas de utilidad de las características podemos encontrar la ganancia de información, el gain ratio, el índice de Gini, ReliefF, etc…

Los algoritmos de selección de características de la tesis de Antonio Arauzo Azofra, se encuentran implementados en el lenguaje de programación Python. Sin embargo, no se pueden acceder fácilmente a ellos desde R, por lo que serán correctamente adaptados al lenguaje de programación R, para que puedan ser añadidos al paquete R que publicaremos.

Por último, además de adaptar algunos de los algoritmos que hemos encontrado mediante las diferentes fuentes, esperamos poder implementar en R algunos métodos adicionales.

# 4.- Descripción del proyecto

Evaluaremos la diversidad de métodos de selección de características, para determinar cuáles son los más adecuadas para un problema dado. Podremos evaluar la eficiencia de los métodos de selección de características, aplicándolos sobre varios conjuntos de datos. Sin embargo, es casi imposible encontrar un algoritmo que sea mejor que los demás para todos los problemas, por lo que buscaremos aquellos que funcionen mejor en los problemas que solemos encontrar, o que funcionen mejor en una clase concreta de problemas.

Una vez elegidos los métodos de selección de características a implementar, separaremos los métodos de búsqueda, de las medidas a utilizar. Ambas partes serán implementadas en el lenguaje de programación R, y a veces será necesario integrarlo con C++. El entorno de programación que utilizaré será R studio.

Los conjuntos de datos que utilizaremos para evaluar el rendimiento de los métodos implementados, podrán ser artificiales (creados a partir de un problema conocido) o reales (datos experimentales obtenidos para resolver un problema real). Un buen repositorio del que usaremos una gran y variada gama de conjuntos de datos, será UCI.

Para medir el rendimiento de un algoritmo de selección características, usaremos algoritmos de aprendizaje conocidos como: naive Bayes, kNN, redes neuronales artificiales, SVM. Evaluando su rendimiento con las características seleccionadas tendremos una medida del rendimiento del algoritmo que ha seleccionado las características.

Tras la correcta implementación de los métodos de búsqueda y las medidas, para conformar los algoritmos de selección de características, diseñaremos una interfaz de usuario que sea fácil de usar e intuitiva, para alguien habituado a un entorno del lenguaje de programación R. La interfaz desarrollada será lo más similar posible a las utilizadas en los diferentes algoritmos de selección de características escritos en R, disponibles en los paquetes del repositorio de CRAN, para que cualquier persona habituada a este entorno pueda utilizar fácilmente nuestros algoritmos.

Una vez dispongamos de los algoritmos listos para utilizarlos, se llevarán a cabo una serie de tests (algunos serán desarrollados e integrados en el paquete para usarse como test automático que facilite la ampliación del paquete sin errores), para comprobar el correcto funcionamiento de los métodos implementados, con las diferentes medidas.

Una vez los algoritmos de selección de características funcionen de forma correcta, se dispondrá a formar el paquete compuesto por los diferentes algoritmos de selección de características, cumpliendo las reglas y los estándares dispuestos por CRAN. Una vez CRAN acepte el paquete realizado, será publicado para su uso y posible mejora.

Por último, realizaremos una experimentación básica comparando los diferentes algoritmos implementados, para analizar los resultados obtenidos.

# 5.- Recursos

Hardware utilizado:

Para la ejecución de los algoritmos de selección de características implementados, se dispondrá de un Cluster en la sala de proyectos LV2P130 con las siguientes características:

* 1 nodo servidor con:
  + 2 procesadores 2.0GHz con 4 cores y RAM de 8Gb.
  + 2 discos duros de 1Tb.
* 7 nodos de cálculo con:
  + 2 procesadores 2.0GHz con 4 cores y RAM de 8Gb.
  + 1 disco duro de 64Gb.

Además cuento con mi portátil particular, para el desarrollo y prueba de los algoritmos programados. Las características del portátil, son las siguientes:

* + 1 procesador Celeron de 2 GHz con 2 cores y RAM de 4Gb.
  + Disco duro de 300Gb.

Software utilizado:

* Se trabajará tanto en el sistema operativo Windows como en Linux, para garantizar que el resultado sea multiplataforma.
* R studio, como entorno de desarrollo.
* Diversos paquetes de R disponibles en CRAN.
* GitHub, como repositorio en el que se alojará el proyecto, de principal importancia para ir modificando el código fuente que iremos implementando.
* Asana, para la organización de tareas y tiempos del trabajo.

# 6.- Plan de Trabajo.

Las principales fases de desarrollo del proyecto en las que se irán consiguiendo los objetivos establecidos, y la distribución temporal de cada una de las tareas, son las siguientes:

|  |  |
| --- | --- |
| ***Fases de desarrollo*** | ***Distribución temporal (horas)*** |
| **Aprendizaje de las tecnologías a usar en el proyecto** | 30 |
| * Lenguaje de programación R | 20 |
| * Paquete Rcpp (API para integrar C++ con R) | 5 |
| * GitHub | 5 |
| **Investigación y selección de algoritmos de selección de características** | 40 |
| * Estudio de los métodos de búsqueda | 15 |
| * Estudio de las medidas | 15 |
| * Selección de los métodos de búsqueda y de las medidas más interesantes | 10 |
| **Implementación y diseño** | 115 |
| * Implementación | 85 |
| * Métodos de búsqueda | 30 |
| * Medidas | 40 |
| * Algoritmos de selección de características | 15 |
| * Diseño | 30 |
| * Paquete R | 10 |
| * Estudio de las normas y estándares que se deben cumplir de acuerdo a CRAN | 10 |
| * Interfaz para probar los algoritmos de selección de características | 10 |
| **Depuración y pruebas experimentales** | 45 |
| * Realización de tests de corrección de las implementaciones | 15 |
| * Realización de pruebas experimentales con cada uno de los algoritmos implementados | 15 |
| * Análisis y evaluación de los resultados obtenidos en los diferentes experimentos realizados | 15 |
| **Publicación del paquete elaborado en CRAN** | 10 |
| **Documentación** | 60 |
| * Memoria técnica | 45 |
| * Manual de uso | 10 |
| * Manual de código | 5 |
| **Total** | 300 |

# Bibliografía

[1] Repositorio de paquetes de aprendizaje automático en R. Disponible vía Web en: https://cran.r-project.org/web/packages/available\_packages\_by\_date.html

[2] Antonio Araúzo Azofra. Un sistema inteligente para selección de características en clasificación. Tesis doctoral (2006). Disponible en: http://ax5.com/wp-content/uploads/seleccion\_de\_caracteristicas\_en\_clasificacion.pdf

[3] Tutorial “Using Rcpp with RStudio”: https://support.rstudio.com/hc/en-us/articles/200486088-Using-Rcpp-with-RStudio

[4] Tutorial de R, disponible vía Web en: http://adv-r.had.co.nz/

[5] Guías básicas para trabajar con Github, disponibles en: http://rogerdudler.github.io/git-guide/index.es.html http://conociendogithub.readthedocs.org/en/latest/data/dinamica-de-uso/#colaborar-en-un-proyecto-ajeno