UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

Semisupervised learning and instance selection methods

Presentado por Daniel Puente Ramírez
en Universidad de Burgos — 30 de diciembre de 2021
Tutor: Álvar Arnaiz González

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

D. Álvar Arnaiz-González, profesor del Departamento de Ingeniería Informática, Área de Lenguajes y Sistemas Informáticos.

Expone:

Que el alumno D. Daniel Puente Ramírez, con DNI dni, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado título de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 30 de diciembre de 2021

Vº. Bº. del Tutor:

D. Álvar Arnaiz-González

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

Índice general

Índice de figuras

Índice de tablas

Introducción

Descripción del contenido del trabajo y del estructura de la memoria y del resto de materiales entregados.

Objetivos del proyecto

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.

Conceptos teóricos

El proyecto tiene una relación directa con la minería de datos y los conceptos que lo rodean.

3.1. Aprendizaje en *machine learning*

En [27] se define *machine learning* como una rama dentro del campo de la Inteligencia Artificial que proporciona a los sistemas la capacidad de aprender y mejorar de manera automática, a partir de la experiencia. Estos sistemas transforman los datos en información, y con esta información pueden tomar decisiones. Este tipo de modelos se crean a base del uso masivo de datos. Cuando se dispone de los datos suficientes para entrenar un modelo comienza el proceso de aprendizaje. El objetivo de este aprendizaje es descubrir patrones ocultos en los datos. En muchas ocasiones el resultado del aprendizaje, el modelo, es una función que dadas unos datos de entrada clasifica o predice correctamente una salida. Como se puede ver en la Figura 3.1 el aprendizaje automático, *machine learning*, posee diferentes aproximaciones, cada una de ellas con una aproximación diferente en cuanto al uso de instancias etiquetadas.

Aprendizaje supervisado

El aprendizaje automático puede ser resumido como aprender de ejemplos. Al programa se le proporcionan dos conjuntos de datos, uno de entrenamiento y otro de validación [20] El objetivo es simple, debe de «aprender» en función del conjunto de datos etiquetado proporcionado como

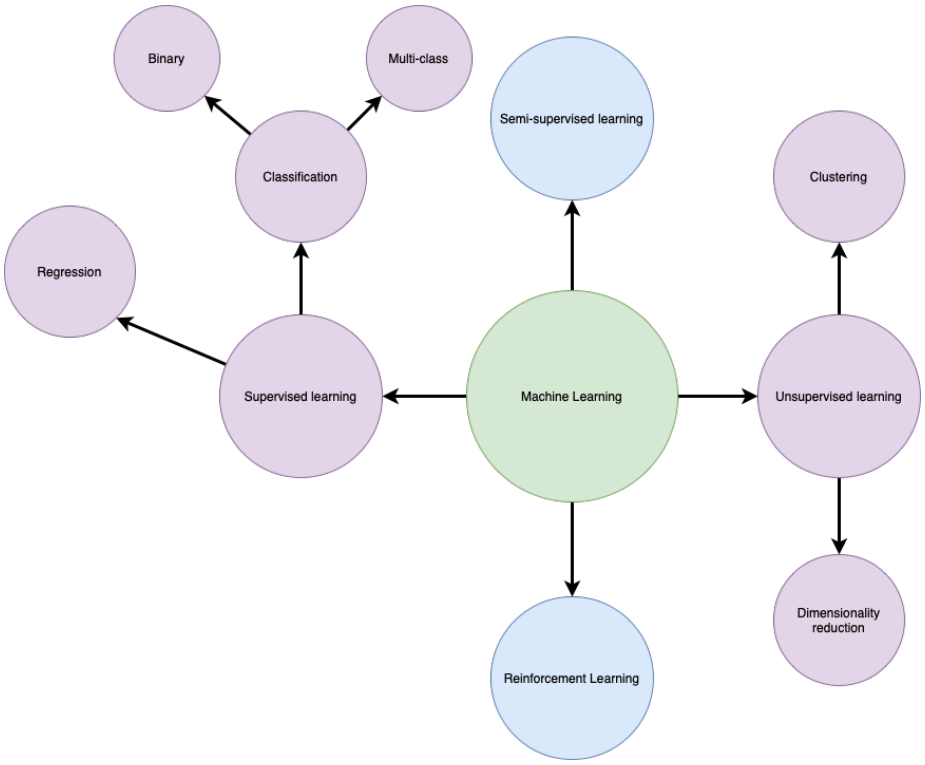

Figura 3.1: *Machine learning overview* [28]

entrenamiento para posteriormente identificar las correspondientes etiqueta/s de cada instancia del conjunto de validación con la mayor precisión posible.

Dependiendo del tipo de etiqueta, en el aprendizaje supervisado hay dos modelos [17]

1. **Modelos de clasificación.** Producen como salida una etiqueta discreta, i.e. una etiqueta dentro de un conjunto finito de etiquetas, habitualmente suelen ser o binarias $[0, 1]$, $[sí, no]$... o multi-etiqueta, donde por ejemplo los valores pueden variar $[0...n]$, i.e. no tienen que ser estrictamente numéricas, pudiendo ser por ejemplo $\{coche, moto, barco\}$. En los modelos de clasificación multi-etiqueta es habitual que el clasificador trabaje con selección de varias etiquetas para la misma muestra, no estando restringido a una única.

   Entre los algoritmos de clasificación más frecuentes encontramos:

   - Regresión logística.

- *Support Vector Machine, SVM.*
- Redes neuronales.
- Clasificador Naïve Bayes.
- Árbol de decisión.
- Análisis discriminante.
- K vecinos más cercanos, *KNN.*
- Clasificación con ensembles.

2. **Modelos de regresión.** Producen como salida un valor real, numérico. Suelen ser soluciones continuas. De igual manera si se quieren obtener varios resultados de una muestra, se utiliza la multi-regresión.

   Entre los algoritmos de regresión más frecuentes encontramos:

   - Regresión lineal.
   - Regresión no lineal.
   - Modelo lineal generalizado.
   - Árbol de decisión.
   - Redes neuronales.
   - Regresión con procesos gaussianos.
   - Regresión con *support vector machines.*
   - Regresión con ensembles.

Aprendizaje no supervisado

En la Sección 3.1 se comenta que, los modelos para que «aprendan» los patrones que se encuentran en los conjuntos de datos, necesitan tener un conjunto de datos etiquetado correctamente para extraer la información de ese conjunto. Pero en los problemas del mundo real no siempre se tienen infinidad de datos disponibles etiquetados correctamente, o simplemente es un proceso muy laborioso y costoso económicamente.

Para solventar este problema se cuenta con el aprendizaje no supervisado [3], mediante esta técnica no es necesario proporcionar al modelo datos etiquetados. Por definición, el algoritmo encargado de entrenar el modelo «aprenderá» los datos sin conocimiento previo. Para ello el modelo se basará en los datos que tiene disponibles y en la codificación del algoritmo para descubrir los patrones que se encuentren en los datos.

Debido a la forma de trabajar del aprendizaje no supervisado, desde el primer momento en el que el algoritmo tiene los datos comienza a reportar salidas, describiendo la información y categorizando lo que encuentra en los datos.

Principalmente existen dos técnicas de aprendizaje no supervisado.

1. ***Clustering*** [22] Proceso por el cual se dividen los datos no clasificados en grupos aparentemente similares. Cuando se identifican datos con algún parecido entre sí, son agrupados. Permite clasificar e identificar atributos únicos de los datos con los que clasificarlos.

   Un proceso habitual de agrupamiento es el uso de *K-means*, $K \in \mathbb{R}$, donde se indica en K cuántos *clusters* o grupos se han de hacer con los datos.

   Con los datos agrupados el proceso de análisis de éstos puede comenzar. En ocasiones si el número de grupos detectados es muy alto, se pueden encontrar grupos o *clusters* irrelevantes, permitiendo a los científicos de datos eliminar esos datos que los forman, reduciendo la dimensionalidad.

2. **Reducción de la dimensionalidad.** La clasificación en el aprendizaje automático se basa en atributos o características que tienen los datos, permitiendo su clasificación, valga la redundancia. Cuando los conjuntos de datos poseen múltiples características, más difícil resulta su clasificación. Es por ello que resulta útil identificar aquellos atributos que están fuertemente interrelacinados entre sí para eliminar todos menos un atributo, reduciendo la dimensionalidad [21]

Aprendizaje semi-supervisado

Semi-Supervised Learning según [34], se define como una forma de entrenamiento de modelos el cual usa tanto datos etiquetados como no etiquetados, i.e. si no sería un aprendizaje supervisado, Sección 3.1, o no supervisado, Sección 3.1.

El uso de aprendizaje semi-supervisado se caracteriza por ser más barato que el supervisado, ya que este último necesita que todo el conjunto de datos que va a utilizar para aprender esté etiquetado, y ese proceso es largo y costoso. Luego, obtiene mejores resultados en menor tiempo que el aprendizaje no supervisado. Conseguir datos sin etiquetar es una tarea

muy sencilla, mientras que conseguir conjuntos de datos etiquetados es un proceso complejo y actualmente no hay «de todo».

Para que el aprendizaje sea fructuoso requiere que las instancias se encuentren inter-relacionadas entre sí por alguna de sus características [18] indica las siguientes suposiciones que se dan en el aprendizaje semi-supervisado.

1. **Continuidad.** Se asume que los objetos cercanos entre sí se encontrarán en el mismo *cluster* o grupo de etiquetas.
2. ***Clustering.*** Las instancias son divididas en diferentes grupos discretos, compartiendo todos los elementos de un *cluster* la misma etiqueta.
3. ***Manifold*** o colectores. Se emplea el uso de distancias y funciones de densidad de forma que las instancias se encuentran en colectores con menos dimensiones que el espacio de entrada.

Dentro de las *best practices* en *semi-supervised learning* se encuentran el uso de diferentes modelos de redes neuronales para el entrenamiento [29]

3.2. Algoritmos en el aprendizaje semi-sepervisado

A continuación se van a presentar los algoritmos tratados en este trabajo.

Self-training

Los métodos de auto-entrenamiento (*self-training*), son uno de los métodos más sencillos de pseudo-etiquetado que existen Triguero *et al.* [30] Se encuentran basados en un único clasificador el cual utiliza aprendizaje supervisado, el clasificador se encuentra siendo entrenado constantemente con datos etiquetados conocidos y por los que se van conociendo a medida que pasan las iteraciones. Es decir, en el inicio el clasificador es entrenado con los datos etiquetados que se conocen, con ese clasificador se obtienen nuevas instancias etiquetadas y las mejores (con mayor *confidence level*) son añadidas al conjunto de datos etiquetado para volver a entrenar el clasificador [11]

Yarowsky [33] en 1995 propuso la primera versión de *Self-training*, desde entonces numerosas aproximaciones han sido realizadas, modificando la utilización del conjunto de datos etiquetado, los nuevos datos, etcétera. El diseño que se le puede dar y los campos de aplicación del mismo son muy variados.

El procedimiento de selección de qué datos son pseudo-etiquetados es de vital importancia, puesto que determinará qué datos acaban en el conjunto de datos etiquetado de sucesivas iteraciones. Siendo este un proceso iterativo y no incremental, ya que la probabilidad de etiquetado de los datos es re-calculada en cada iteración. De no serlo sería una apoximación a *expectation-maximization* [9]

Co-Training

Blum [4] en 1998 propuso el *Co-Training* para conjuntos de datos compuestos por datos etiquetados y no etiquetados. Bajo la presunción de que con unos pocos datos etiquetados y diferentes clases que aportan información, se pueden entrenar dos algoritmos de aprendizaje por separado para posteriormente añadir al conjunto de datos etiquetados aquellas predicciones con mayor *confidence level*.

Las dos características del problema mencionadas anteriormente, disponibilidad de datos etiquetados y no etiquetados, y la disponibilidad de dos «tipos» diferentes de conocimiento sobre los ejemplo; aproximan a la siguiente estrategia de aprendizaje. Se desea encontrar los predictores débiles basados en cada tipo de información utilizando un pequeño conjunto inicial de instancias etiquetadas, seguidamente, utilizando los datos no etiquetados se intenta hacer un *bootstrap* a partir de esos «malos» predictores. Este tipo de *bootstrapping* es el denominado *Co-Training*, y posee una estrecha relación con el *bootstraping* a partir de datos incompletos en el marco de la maximización de expectativas [13, 25]

Algorithm 1 *Co-Training.*

Require: Conjunto de entrenamiento $L\{(x_i, y_i)\}_{i=1}^{l}$ y $U\{x_j\}_{j=l+1}^{l+u}$ de datos etiquetados y no etiquetados, respectivamente

Require: p, n, k, u, datos positivos a seleccionar, los negativos, iteraciones, tamaño *pool* inicial

Ensure: Conjunto etiquetado $S \subset \{L, U\}$

procedure CO-TRAINING(L, U, p, n, k, u)
  $U' \leftarrow U[u]$▷ u instancias aleatorias de U
  for k **do**
    Usar L para entrenar un clasificador h_1 utilizando x
    Usar L para entrenar un clasificador h_2 utilizando y
    h_1 clasifica U'
    h_2 clasifica U'
    Seleccionar las p y n instancias con mayor *confidence level* de cada clasificador
    Añadirlas a L y eliminarlas de U'
    Elegir de forma aleatoria $2p + 2n$ instancias de U y añadirlas a U'
  end for
end procedure

Tri-Training

Zhou [36] en 2005 propuso una modificación sobre el algoritmo de *Co-Training* de Blum [4] el cual requería de dos «vistas significativas». Dasgupta *et al.* [8] demostraron que cuando los requerimientos del conjunto de datos se cumplían, se podían producir un menor número de errores de generalización al maximizar la clasificación de los prototipos sin etiquetar, aunque no es aplicable a cualquier conjunto de datos.

El *Tri-Training* se define como una nueva aproximación de *Co-Training*. *Tri-Training* no necesita varias «vistas significas» de los datos, tampoco requiere el empleo de múltiples algoritmos de aprendizaje supervisado cuyas hipótesis dividen el espacio de instancias en un conjunto de clases de equivalencia. El *Tri-Training* por tanto emplea tres clasificadores, a diferencia de los dos utilizados anteriormente, esta opción resuelve la dificultad de determinar cómo etiquetar las instancias no etiquetadas y generar la hipótesis final, lo que mejora enormemente la eficiencia del algoritmo.

Junto con la capacidad de generalización resultante de la combinación de los tres clasificadores.

Los tres clasificadores, h_1, h_2 y h_3, son entrenados inicialmente con todo el conjunto de datos etiquetado. Seguidamente, a cualquier instancia no etiquetada, se la podrá asignar una etiqueta siempre y cuando hay al menos dos clasificadores de acuerdo con la asignación, i.e. $label(x) = h_i(x) = h_j(x)$, ver algoritmo 2. Puede darse el caso de que dos clasificadores acierten en la predicción y la etiqueta sea considerada correcta, pero para el tercer clasificador sea ruido, incluso en el peor caso, el aumento del ruido en el proceso de clasificación puede mitigarse si el número de instancias recién etiquetadas es significativo (bajo condiciones específicas) [36]

Debido a que *Tri-Training* no asume la existencia de clases «redundantes», se necesita un cierto grado de diversidad en los clasificadores. Esta diversidad es alcanzada mediante la manipulación del conjunto de datos etiquetado. Los clasificadores iniciales son entrenados con los datos generados mediante *bootstrap*[1] del conjunto de datos etiquetados original. Estos clasificadores son depurados en el proceso iterativo del algoritmo, produciendo la hipótesis final mediante mayoría simple.

Dado que *Tri-Training* no impone ninguna restricción al algoritmo de aprendizaje supervisado ni emplea un proceso de validación cruzada que requiera mucho tiempo de cómputo, tanto su aplicabilidad como su eficiencia demuestran ser mejores que otras versiones de *Co-Training*.

[1] En el campo de la estadística, se define como un método el cual consiste en la extracción de datos de muestra repetidamente con reemplazo de un conjunto de datos, con el fin de estimar un parámetro de la población.

Algorithm 2 *Tri-Training.*

Require: Conjunto de entrenamiento L y U de datos etiquetados y no etiquetados, respectivamente
Require: *Learn*: algoritmo de aprendizaje
Ensure: Conjunto etiquetado $S \subset \{L, U\}$

procedure TRI-TRAINING($L, U, Learn$)
  for $i \in \{1.,3\}$ **do**
    $S_i \leftarrow BootstrapSample(L)$
    $h_i \leftarrow Learn(S_i)$
    $e'_i \leftarrow 0{,}5$
    $l'_i \leftarrow 0$
  end for
  repeat
    for $i \in \{1.,3\}$ **do**
      $L_i \leftarrow \emptyset$
      $update_i \leftarrow$ **False**
      $e_i \leftarrow MeasureError\left(h_j \& h_k\right)\ (j,k \neq i)$
      if $e_i < e'_i$ **then**
        for all $x \in U$ **do**
          if $h_j\left(x\right) = h_k\left(x\right)\ (j,k \neq i)$ **then**
            $L_i \leftarrow L_i \cup \{(x, h_j\left(x\right))\}$
          end if
        end for
        if $l'_i = 0$ **then**
          $l'_i \leftarrow \left\lfloor \frac{e_i}{e'_i - e_i} + 1 \right\rfloor$
        end if
        if $l'_i < |L_i|$ **then**
          $update_i \leftarrow$ **True**
        else if $l'_i > \frac{e_i}{e'_i - e_i}$ **then**
          $L_i \leftarrow Subsample\left(L_i, \left\lceil \frac{e'_i l'_i}{e_i} - 1 \right\rceil\right)$
          $update_i \leftarrow$ **True**
        end if
      end if
    end for
    for $i \in \{1.,3\}$ **do**
      if $update_i =$ **True then**
        $h_i \leftarrow Learn\left(L \cup L_i\right)$
        $e'_i \leftarrow e_i$
        $l'_i \leftarrow |L_i|$
      end if
    end for
  until ningún $h_i\,(i \in \{1.,3\})$ cambie
end procedure

Democratic Co-Training

Zhou [35] en 2004 presentó el algoritmo *Democratic Co-Learning*. El algoritmo a diferencia de sus «homónimos», trabaja con múltiples algoritmos de aprendizaje supervisado, en lugar de múltiples clases significativas, permitiendo que se etiqueten nuevas instancias entre ellos. Debido a que diferentes algoritmos de aprendizaje poseen diferentes sesgos, seleccionar la clase más votada por la mayoría produce mejores predicciones.

El algoritmo es por tanto enfocado para el uso en casos donde:

- Solo se poseen unas pocas muestras etiquetadas
- Existe un conjunto de datos de tamaño muy superior, sin etiquetar
- No existen dos conjuntos de atributos independientes y redundantes

Es por ello que *Democratic Co-Learning* utiliza un conjunto de datos etiquetados, L, uno de no etiquetados, U, y $A_1, \ldots, A_n$ para $n \geq 3$, siendo n los algoritmos de aprendizaje supervisado, ver algoritmo 3. El algoritmo comienza entrenando los n clasificadores sobre el conjunto de datos etiquetado L; y para cada instancia x del conjunto de no etiquetados U, cada clasificador predice una etiqueta $c_i \in \mathcal{C} = \{c_1, c_2, \ldots, c_r\}$. Siendo c_k la predicción mayoritaria. Posteriormente los n clasificadores son re-entrenados con los nuevos datos etiquetados (añadidos a los que ya teníamos), este proceso se realiza de manera iterativa hasta que no se seleccionen más datos para realizar el etiquetado. Para la selección final se realizar un voto mayoritario ponderado entre los n clasificadores.

Una instancia x nunca será etiquetada por la decisión única de un clasificador, a menos que la mayoría de los clasificadores estén de acuerdo. Además se requiere que la suma de los valores medios de confianza de los clasificadores del grupo mayoritario sea mayor, que la suma de los valores medios de confianza de los clasificadores de los grupos minoritarios, siendo la confianza media de un clasificador $(l + h)/2$ para l y h definidas por el intervalo de confianza del 95 % $[l, h]$.

Finalmente el algoritmo, ver 4, finaliza con la combinación de todas las hipótesis generadas para retornar la hipótesis final. Para realizar el cálculo utiliza un sistema de votación mayoritaria de entre las posibles clases, para hilar un poco más fino, se considera también para cada clasificador su valor de confianza de la predicción.

Algorithm 3 *Democratic Co-Learning.*

Require: Conjunto de entrenamiento L y U de datos etiquetados y no etiquetados, respectivamente

Require: $A_1, \ldots, A_n$ los n algoritmos de aprendizaje supervisado

Ensure: Conjunto etiquetado $S \subset \{L, U\}$

procedure DEMOCRATIC CO-LEARNING(L, U)
  for $i = i, \ldots, n$ **do**
    $L_i \leftarrow L$
    $e_i \leftarrow 0$
  end for
  repeat
    for $i = 1, \ldots, n$ **do**
      Calcular las hip H_i con los datos L_i para cada clasificador A_i
    end for
    for all $x \in U$ **do**
      for $j = 1, \ldots, r$ **do** ▷ Posibles etiquetas
        $c_j \leftarrow |\{H_i | H_i(x) = j\}|$
      end for
      $k \leftarrow argmax_j\{c_j\}$
    end for
    for $i = 1, \ldots, n$ **do** ▷ x propuestos para etiquetar
      Utilizar L para calcular el int. de conf. 95 % $[l_i, h_i]$ para H_i
      $w_i \leftarrow (l_i + h_i)/2$
      for $i = 1, \ldots, n$ **do**
        $L'_i = \emptyset$
      end for
      if $\sum_{H_j(x)=c_k} w_j > max_{c'_k \neq c_k} \sum_{H_j(x)=c'_k} w_j$ **then**
        $L'_i \leftarrow L'_i \cup \{(x, c_k)\}$, $\forall i$ tal que $H_i(x) \neq c_k$
      end if
    end for
  ▷ Comprobar si añadir L'_i a L mejora la precisión
    for $i = 1, \ldots, n$ **do**
      $q_i \leftarrow |L_i| \left(1 - 2\left(\frac{e_i}{|L_i|}\right)\right)^2$ ▷ est. del error
      $e'_i \leftarrow \left(1 - \frac{\sum_{i=1}^{d} l_i}{d}\right) |L'_i|$ ▷ est. del nuevo error
      $q'_i \leftarrow |L_i \cup L'_i| \left(1 - \frac{2\left(e_i + e'_i\right)}{\left|L_i \cup L'_i\right|}\right)^2$ ▷ si L'_i es añadida
    end for
    if $q'_i > q_i$ **then**
      $e_i \leftarrow e_i + e'_i$
    end if
  until Ningún $L_1, \ldots, L_n$ cambie

Algorithm 4 *Democratic Co-Learning.*

for $i = 1, \ldots, n$ **do**
  Calcular el int. de conf. 95 % $[l_i, h_i]$ para H_i usando L
  $w_i \leftarrow (l_i + h_i)/2$
end for
for all $x \in (L \cup U)$ **do**
  for $i = 1, \ldots, n$ **do**
    if $H_i(x)$ predice c_j y $w_i > 0{,}5$ **then**
      Asociar H_i al grupo G_j
    end if
  end for
  for $j = 1, \ldots, r$ **do**
    $\overline{C}_{G_j} \leftarrow \frac{|G_j|+0{,}5}{|G_j|+1} \times \frac{\sum_{H_i \in G_j} w_i}{|G_j|}$
  end for
end for
Predicciones H con Gk para $k = argmax_j \left(\overline{C}_{G_j}\right)$
return H
end procedure

3.3. Minería de datos

Según IBM [10], podemos definir la minería de datos, o descubrimiento de conocimiento en los datos *Knowledge Discovery in Databases (KDD)*, como el proceso de descubrir patrones y otra información a partir de grandes conjuntos de datos.

Las técnicas de minería de datos principales se pueden dividir en función de sus propósitos principales.

1. Descripción del conjunto de datos objetivo.
2. Predicción de resultados mediante el uso de algoritmos de aprendizaje automático.

Proceso de minería de datos

El proceso de minería de datos comprende varios pasos como crear, probar y trabajar con los modelos de minería. Comienza con la recogida de los datos que van a ser tratados, y finaliza con la visualización de la información extraída de éstos. Los científicos de datos describen los datos a

través de sus observaciones de patrones, asociaciones y correlaciones. A su vez se pueden clasificar y agrupar los datos utilizando métodos de clasificación y regresión.

Uno de los marcos de referencia más importantes en el proceso de minado de datos es CRISP-DM, *Cross Industry Standard Process for Data Mining*. Desarrollado por un consorcio de empresas involucradas en la minería de datos [6]

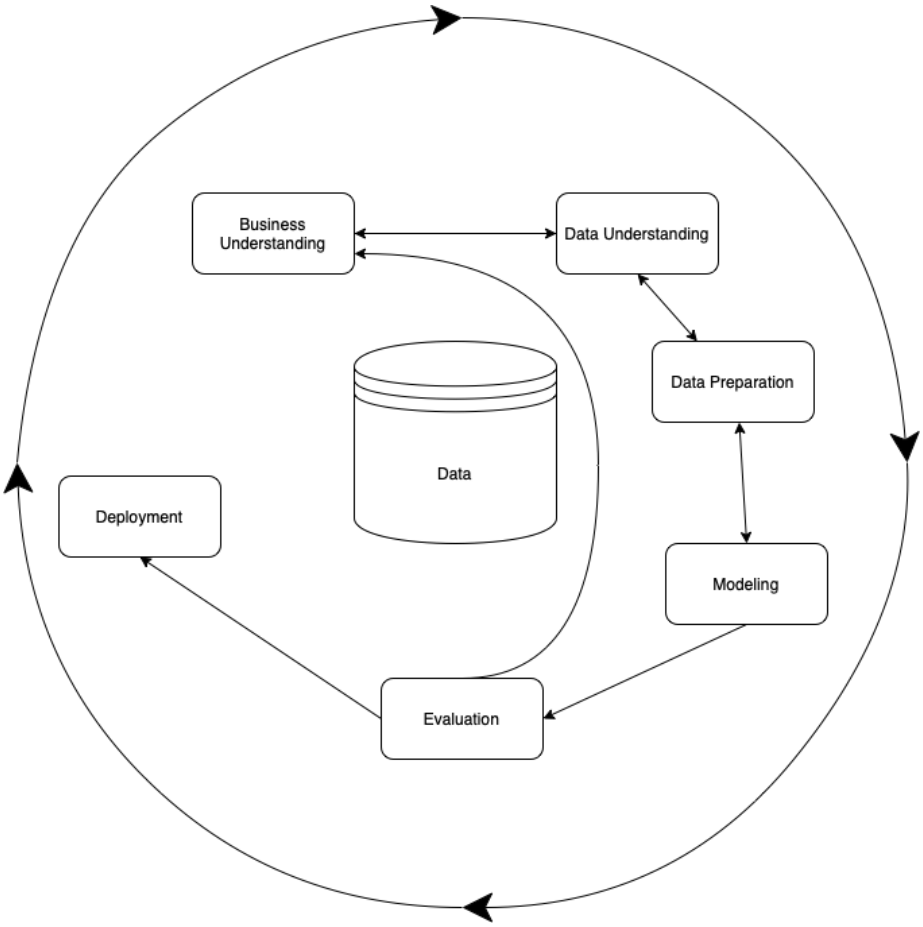

Figura 3.2: Enfoque CRISP de la minería de datos [19]

En [19] se divide el proceso de la minería de datos en 5 etapas o pasos principales: establecimiento de los objetivos y comprensión del problema, recopilación y preparación de los datos, desarrollo del modelo, aplicación del modelo y la evaluación de los resultados y despliegue en producción. Ver Figura 3.3.

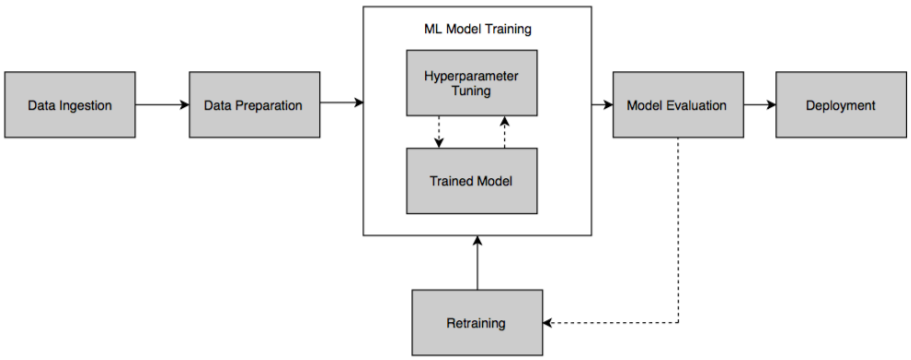

Figura 3.3: *Machine Learning Pipeline* [1]

1. **Establecer los objetivos y comprensión del problema.** La primera etapa puede resultar la más complicada del proceso. Todas las partes interesadas deben de estar presentes y de acuerdo en la definición del problema que se va tratar, esto incluye tanto a los científicos de datos como las terceras partes involucradas o interesadas. Este procedimiento ayuda a la formulación de las preguntas de los datos y los parámetros a utilizar en el proyecto. Si se trata de un proyecto empresarial, se debe hacer un estudio o investigación adicional para comprender el contexto de la empresa.

2. **Preparación de los datos.** Con el alcance del problema definido ya se puede comenzar a identificar qué conjunto de datos será el más efectivo o representativo con el fin de comenzar a dar respuesta a las preguntas formuladas en el proceso anterior.

   Una vez se dispone de todos los datos recogidos comienza el proceso de pre-procesado de los mismos. Este proceso se basa en la limpieza de los datos con el fin de eliminar cualquier posible ruido, entendiéndose por ruido los datos duplicados, los valores perdidos y aquellos atípicos; aquellos que puedan causar problemas a la resolución del problema o generen incertidumbre. En determinados conjuntos de datos se puede hacer una reducción de dimensiones. Consiste en la reducción del número de dimensiones que poseen las instancias recogidas, con el fin de eliminar aquellas que no sean realmente representativas o significativas, este proceso reduce la complejidad de los cálculos posteriores. Por contrapartida hay que conocer cuáles serán los predictores con mayor relevancia en el problema para garantizar una precisión «óptima» del modelo.

3. **Desarrollo del modelo.** Según [19] el modelo es la representación abstracta de los datos y sus relaciones en un conjunto de datos concreto. Actualmente existen cientos de algoritmos que se pueden utilizar, habitualmente proceden de campos como la ciencia de datos, *machine learning*, o la estadística. Se debe tener el conocimiento suficiente para entender como funciona el algoritmo para poder configurar correctamente los parámetros que este va a utilizar en base a los datos y el problema de negocio que estamos resolviendo.

   Los modelos en función de como resuelvan el problema que se les presenta se pueden clasificar en:

   a) Regresión.

   b) Análisis de asociación.

   c) *Clustering.*

   d) Detección de anomalías.

   El modelo debe ser creado con especial cuidado para evitar el *overfitting*, i.e. el modelo memoriza el conjunto de entrenamiento y no tendrá un rendimiento correcto una vez desplegado en producción. Se desea que el modelo sea lo más general posible de cara a *aprender* de los datos del conjunto de entrenamiento.

4. **Aplicación del modelo.** El momento de la aplicación del modelo es cuando de verdad se comprueba si realmente el modelo está listo para pasar al siguiente punto, en otras palabras, si es apto para ser desplegado en producción. Para ello se tienen en cuenta métricas como la calidad del modelo ante el problema, su tiempo de respuesta, etc.

5. **Evaluación de los resultados y despliegue en producción.** Es habitual que los parámetros con los que el modelo fue entrenado con el paso del tiempo dejen de ser los más interesantes, pudiendo ser comprobado el error proporcionado por el modelo con los datos de prueba. Cuando ese error sea excesivo o fuera de un margen dado se deberá de volver a entrenar el modelo, comprobar, y desplegar. De esta forma se puede comprobar como el ciclo de vida del modelo es circular.

El proceso aplicado en la minería de datos proporciona un marco de trabajo mediante el cual se permite extraer información aparentemente no trivial de grandes conjuntos de datos. Es un campo de aprendizaje constante,

tanto el aplicar los conocimientos del analista para reducir las dimensiones del conjunto de datos, como una vez que se ha entrenado el modelo y puesto en producción, aprender los puntos fuertes de este y el por qué de éstos [6]

Técnicas utilizadas en la minería de datos

A continuación se presentan una serie de técnicas utilizadas en función de la naturaleza de las instancias predictoras, y de la variable de salida. Si la variable o clase de salida es continua o categórica, nos encontramos con modelos de aprendizaje supervisado, mientras que si no existe variable o clase de salida, nos encontramos con modelos de aprendizaje no supervisado [24]

1. **Reglas de asociación** (además se trata de una técnica de aprendizaje automático). Se basa en el uso de reglas básicas qutilizadas para localizar relaciones entre instancias en conjuntos de datos de gran tamaño. Para su correcto funcionamiento deben satisfacer el soporte (nivel) mínimo especificado por el usuario y la confianza o grado de satisfacibilidad especificada en tiempo constante.

   En determinadas ocasiones la generación de estas reglas es dividida en una serie de pasos:

   - Para encontrar las n instancias más frecuentes, se aplica un umbral mínimo, lo cual establece la información del conjunto de datos.
   - Cuando el nivel mínimo de confianza es aplicable a aquellas instancias encontradas en el paso previo, se transforman en reglas. Este paso es el que más atención requiere.

2. **Redes neuronales.** Principalmente utilizadas en *deep learning*, simulan la interconectividad propia del cerebro humano utilizando capas de nodos. Cada nodo está compuesto por x_n entradas, w_n pesos y un sesgo o umbral, el cual al ser superado activa la neurona, pasando los datos del nodo a la siguiente neurona. El proceso es repetido a lo largo de n iteraciones pasando el mismo conjunto de entrenamiento, conocido como *epochs*.

   Las redes neuronales poseen tres ventajas en el uso de grandes conjuntos de datos: aprendizaje adaptado mediante ejemplos, robustez en el manejo de información redundante e imprecisa, y computación masiva paralela.

3. **Árboles de decisión.** Se pueden definir como particiones secuenciales de un conjunto de datos, maximizando las diferencias a una variable dependiente. Ofrecen una forma concisa de definir grupos que son consistentes en sus atributos pero que varían en términos de la variable dependiente.

   Los árboles de decisión se encuentran compuestos de nodos (variables de entrada), ramas (grupos de variables de entrada), y hojas (valores de la variable de salida). La construcción de los árboles está basada en el principio de *divide and conquer*, haciendo uso de un algoritmo de aprendizaje supervisado, se realizan divisiones sucesivas del espacio multi-variable con el objetivo de maximizar la distancia entre los grupos de cada división, i.e. realizar particiones discriminatorias. El proceso de división finaliza cuando todas las entradas de una rama tienen el mismo valor en el nodo hoja, dando lugar al modelo completo. Cuanto más abajo estén las variables de entrada en el árbol, menos importantes son el la clasificación de la salida.

   Para evitar el *overfitting* del modelo, el árbol puede podarse eliminando las ramas con pocas instancias, o donde aquellas instancias sean poco representativas [24]

4. **k-vecinos más cercanos. KNN (*k-nearest neighbors*)** [14, 15, 24] Las técnicas k-NN se basan en el concepto de similaridad. Permite la construcción de un método de clasificación sin hacer suposiciones sobre la forma de la función que relaciona la variable dependiente con las variables independientes.

   El objetivo es identificar de forma dinámica las k instancias en los datos de entrenamiento que son similares (vecinas) a la instancia que se quiere clasificar. La clasificación de una instancia viene dada por la observación de la clase de la vecindad, para ello se basa en los atributos de las variables. En otras palabras, cuenta el número de instancias para cada clase en la vecindad y asigna a la instancia en cuestión aquella clase que sea mayoritaria en la vecindad [7]

   Asume que todas las instancias corresponden a puntos en un espacio n-dimensional. Pudiendo ser utilizado tanto en problemas de clasificación como de regresión.

Figura 3.4: Proceso de selección de instancias.

3.4. Técnicas de selección de instancias

Dentro de los conjuntos de datos nos encontramos con las instancias, también llamadas ejemplos o prototipos, son cada uno de los elementos que componen el *dataset*; en problemas reales de *machine learning* es habitual que se requiera de clasificación automática de estos datos. Este proceso se puede llevar a cabo con algoritmos de aprendizaje supervisado, Sección 3.1, con el objetivo de etiquetar la nueva información. Para poder hacerlo previamente se ha tenido que entrenado el clasificador con un conjunto de entrenamiento, T [23]

En la práctica, cualquier T dado contendrá información útil e información desechable, este último tipo de información — que en realidad son instancias — a parte de ser redundantes producen ruido, pudiendo inducir en una clasificación errónea en el proceso de aprendizaje, y posteriormente tener un modelo que no sea capaz de clasificar correctamente la nueva información.

Es por ello que un pre-procesado o **filtrado** de las instancias pertenecientes al T original es necesario. En la Figura 3.4 se puede consultar de manera gráfica el proceso de selección de instancias. Dado un conjunto de datos de entrenamiento inicial, T, el objetivo será obtener un subconjunto S, tal que $S \subseteq T$ de manera que S no contiene instancias redundantes ni «ruidosas». Además, $Acc(S) \cong Acc(T)$, donde $Acc(X)$ es la precisión, *accuracy* en inglés, del modelo entrenado con el conjunto de datos X.

En función de cómo comienzan a crear el nuevo subconjunto de datos, S, se identifican dos aproximaciones, ascendente y descendente.

- **Ascendente.** El nuevo conjunto de datos comienza estando vacío, $S = \emptyset$, y a medida que se vayan realizando iteraciones del algoritmo correspondiente, se irán añadiendo instancias a S. El principal problema que posee esta aproximación es su sensibilidad al orden, i.e. dada una instancia $x \subset T$, en diferentes iteraciones del mismo algoritmo de selección de instancias sobre el mismo T, puede o no estar en S. Esto se debe a la aleatoriedad con la que se presentan los datos, para asegurar esta aleatoriedad los datos se escogen de manera aleatoria de T, intrínsecamente da una mayor facilidad a las muestras iniciales a estar en S que las finales, ya que puede que ya se encuentren representadas o sean clasificadas como ruido.

  Entre las principales ventajas de esta aproximación destaca el espacio de almacenamiento requerido, puesto que se van guardando instancias y por lo tanto en un inicio es muy pequeño.

- **Descendente.** El nuevo conjunto de datos comienza siendo el conjunto de entrenamiento al completo, $S = T$, y a medida que se vayan realizando las iteraciones del algoritmo correspondiente, se irán eliminando instancias de S. Esta aproximación es mucho más costosa computacionalmente, para cada instancia que debe decidir si eliminar o no debe comprobar todo el subconjunto S, pero en contraposición consigue reducir más que la aproximación ascendente, el conjunto de entrenamiento T.

Junto con esta diferenciación, en función de la aproximación de selección de instancias se pueden distinguir dos agrupaciones.

- ***Wrapper.*** El criterio de selección se basa en la precisión del clasificador. Aquellas instancias que no contribuyen a la mejora del clasificador se quedan fuera de S. Este trabajo está centrado en este criterio de selección.

- ***Filter.*** El criterio de selección utiliza una función, $f(x, y)$, para realizar la selección, no se basa en un clasificador concreto.

En la Tabla 3.1 se aprecian aquellos algoritmos implementados en primera instancia para la reducción de instancias dentro de T, el objetivo

Método	Basado en	Referencia
ENN	Clasificación incorrecta	[32]
CNN	Clasificación incorrecta	[16]
RNN	Clasificación incorrecta	[12]
ICF	Alcance y covertura	[5]
MSS	Fronteras de decisión	[2]

Tabla 3.1: Algunos métodos de selección de instancias.

de todos ellos es que el subconjunto generado, S, sea capaz de clasificar correctamente T en su totalidad prácticamente.

Algoritmos de selección de instancias

Existen multitud de algoritmos a día de hoy que son capaces de reducir el número de instancias de T, en este trabajo se van a comentar los pertenecientes a la Tabla 3.1. Cada uno de ellos tiene sus ventajas y sus desventajas como cabe esperar, en la literatura no apreciamos un "este es mejor que aquel" o similar.

Algoritmo de edición de Wilson

Wilson [32] (1972) publicó la regla del vecino más cercano editado, *ENN*. Los problemas de clasificación de instancias en función de una etiqueta, se caracterizan por:

- Hay una instancia a ser clasificada.
- Existe un S el cual posee instancias con la misma distribución que la instancia a clasificar, pudiendo ser comparables las instancias del conjunto con la que estamos analizando.
- No existe información adicional del conjunto.
- Existe una distancia medible entre instancias.

Con todas estas premisas Wilson propone un algoritmo basado en clasificación incorrecta en función de sus vecinos más cercanos. Cuando

una instancia resulta mal clasificada por sus k vecinos más cercanos, k-NN, esa instancia es descartada. Finalmente obtendremos como resultado un conjunto S con las instancias correctamente clasificadas por sus vecinos.

Suponiendo que sea X un conjunto de N instancias y M posibles clases y, sea k el número de vecinos cercanos,el algoritmo de Wilson se puede formular de la siguiente manera:

Algorithm 5 Algoritmo de edición de Wilson, *ENN*.

Require: Conjunto de entrenamiento, $X = \{(x_1, y_1)...(x_n, y_n)\}$, k vecinos.
Ensure: Conjunto editado $S \subset X$.

```
1: procedure ENN(X, k)
2:     S ← X
3:     for all x ∈ S do
4:         Find x.N_{1...k+1}, the k + 1 nearest neighbors of x ∈ X − {x}
5:         if δ_{k−NN}(x_i) ≠ θ_i then
6:             Remove x of S
7:         end if
8:     end for
9:     return S
10: end procedure
```

El algoritmo de edición de Wilson, ver algoritmo 5 posee una complejidad computacional de $O(n^2)$. Una de las ventajas de este algoritmo es su forma de crear el subconjunto S, ya que al ser descendente las primeras iteraciones serán lentas — dependiendo del tamaño de T lógicamente — pero las finales serán considerablemente más rápidas.

Algoritmo Condensado de Hart

Hart [16] en 1968 propuso la que se considera la primera regla formal de condensador para NN. El algoritmo de condensado de Hart, *CNN — Condensed Nearest Neighbor*. Está basado en técnicas de consistencia y reducción. Sea $X \neq \emptyset$ y $S \subseteq X$, podremos decir que el subconjunto S es consistente respecto al conjunto X si al utilizar a S como conjunto de aprendizaje, se puede clasificar correctamente a todo el conjunto X.

Algorithm 6 Algoritmo Condensado de Hart, *CNN*.
Require: Conjunto de entrenamiento X
Ensure: Conjunto editado $S \subset X$

```
1: procedure CNN(X)
2:     S ← {x_1}
3:     for all x ∈ X do
4:         if x no se clasifica correctamente usando S then
5:             Añadir x a S
6:             Restart
7:         end if
8:     end for
9: end procedure
```

El algoritmo de Hart es una técnica ascendente, a partir de las primeras instancias que se añadan a S se clasificarán y añadirán, o no, a S. Consiste en encontrar entre todas las instancias de T un subconjunto S tal que cada instancia de T sea más cercano a las instancias en S de su misma clase que a las instancias de otras clases, permitiendo utilizar S como conjunto de clasificación de T. Para el correcto funcionamiento se asume que el conjunto T es consistente, no posee dos instancias idénticas con pertenencia a diferentes clases.

El algoritmo propuesto por Hart [16], ver algoritmo 6, posee una complejidad de $O(n^2)$. El conjunto obtenido a partir de T, i.e. S, operando con grandes conjuntos de datos demuestra poseer un tamaño considerablemente menor respecto a T.

Si bien es una técnica utilizada por su efectividad, posee una serie de puntos negativos a su vez.

- Sensibilidad ante el ruido. Un objeto ruidoso no será correctamente clasificado por sus vecinos. Estas muestran no se eliminarán del conjunto solución S, por lo que no desaparecerán.

- S no tiene porqué ser el menor conjunto de T. Diferentes ejecuciones del algoritmo sobre el mismo T pueden dar diferentes conjuntos solución S. Esto se debe al orden aleatorio por el cual se seleccionan las instancias. Por definición del propio algoritmo se asume que no se va a alcanzar

de forma general el subconjunto de tamaño mínimo que cumpla con las características especificadas.

Algoritmo Condensado Reducido

Gates [12] en 1972 propuso el algoritmo del conjunto reducido, basado en las reglas *NN*. El algoritmo propuesto es una modificación del algoritmo *CNN*, ver algoritmo 6. No es una nueva regla de decisión puesto que se sigue eligiendo la clase del vecino más cercano para la clasificación. El algoritmo se basa en un procedimiento para seleccionar el subconjunto T_{CNN}, el cual debe comportarse igual de bien que T_{NN} ante clasificaciones de instancias desconocidas. Como se puede apreciar en el propio algoritmo 7, puede existir una disminución del rendimiento, a cambio se obtiene una mejora de la eficiencia para el algoritmo de clasificación que posteriormente se utilice, tanto en la cantidad de memoria utilizada como en el tiempo de computación.

De igual manera que en *CNN*, posee la problemática de la minimalidad, si bien el conjunto resultante $S \subset T$ será consistente, no se puede asegurar que sea mínimo; y diferentes ejecuciones del algoritmo sobre el mismo conjunto de datos T, pueden obtener diferentes subconjuntos solución S.

Algorithm 7 Algoritmo Condensado Reducido, *RNN*.

Require: Conjunto de entrenamiento X
Ensure: Conjunto editado $S \subset X$

```
1:  procedure RNN(X)
2:      S ← {x_1}
3:      for all x ∈ S do
4:          if x no se clasifica correctamente usando S then
5:              Añadir x a S
6:              Restart
7:          end if
8:      end for
9:      for all x ∈ S do
10:         Remove x de S
11:         if ∃x_i incorrectamente clasificada usando S then
12:             Añadir x a S
13:         end if
14:     end for
15: end procedure
```

Algoritmo *Iterative Case Filtering*

Brighton [5] en 2002 propuso el algoritmo iterativo de filtrado, *ICF*, bajo la premisa de predecir la clase de una instancia con la misma precisión, o mayor si fuera el caso, que el T original. Uno de los objetivos principales del propio algoritmo es mantener en el subconjunto S únicamente aquellas instancias que sean críticas para la decisión.

Algorithm 8 Algoritmo *Iterative Case Filtering*, *ICF*.

Require: Conjunto de entrenamiento X
Ensure: Conjunto editado $S \subset X$

```
1:  procedure ICF(X)
2:      T ← ENN(X, k) ▷ Filtro de ruido en función de la regla k-NN
3:      repeat
4:          for all x ∈ T do
5:              Calcular coverage(x)
6:              Calcular reachable(x)
7:          end for
8:          progress ← false
9:          for all x ∈ T do
10:             if |reachable(x)| > |coverage(x)| then
11:                 Marcar x para eliminar
12:                 progress ← true
13:             end if
14:         end for
15:         for all X ∈ T do
16:             Eliminar x de T
17:         end for
18:     until no progress
19: end procedure
```

ICF puede describirse como un filtro de borrado de instancias, más que de clasificación, con el objetivo de eliminar aquellas instancias que sean superfluas o que aportern ruido a T [5] Al reducir el tamaño de T los tiempos de respuesta para el proceso de clasificación mejorarán, ya que se examinarán un menor número de instancias para realizar la clasificación; por contrapartida al eliminar muestras que se creen que son dañinas para el proceso, se pueden estar eliminando algunas que sean clave y por lo tanto teniendo una degradación de la calidad del clasificador.

ICF se basa en dos categorías para la decisión de si una instancia debe permanecer o no en S, ver algoritmo 8, estas son *Coverage* y *Reachable*. Definidas de la siguiente manera para el caso base $\mathcal{CB} = \{x_1, x_2, ..., x_n\}$.

$$Coverage(x) = \{x' \in \mathcal{CB} : Adaptable(x, x')\}$$
$$Reachable(x) = \{x' \in \mathcal{CB} : Adaptable(x', x)\}$$

Una instancia x podrá estar en el conjunto adaptable de x' si y solo si x es una instancia relevante para la solución de x', i.e. $x \in k - NN(x)$. La problemática surge en el momento en el cual una instancia con clase diferente no permite la correcta clasificación de x', es por ello que el vecindario de x' viene definido por todas las instancias antes de la primera instancia de diferente clase, utilizando *sets* descritos por Wilson y Martinez. La propiedad *Reachable* no está fijada desde el principio del algoritmo, sino que es dinámica siendo fijada cada vez en función de la instancia de clase diferente más cercana. El criterio seguido para realizar la eliminación de una muestras es: si el *set* formado por el $reachable(x)$ es mayor que el $coverage(x)$, i.e. una instancia x es eliminada cuando más instancias pueden resolver x que las que x puede resolver por sí misma.

Algoritmo Subconjunto Selectivo
Ritter [26] 1975 propone un algoritmo capaz de satisfacer la condición de consistencia del algoritmo condensado de Hart, ver 3.4; para ello introduce una condición más "fuerte" de consistencia, el objetivo es encontrar aquellas instancias de un orden independiente.

Un subconjunto S del conjunto de entrenamiento T, es un Subconjunto Selectivo, SS, si SS cumple las siguientes condiciones:

1. El subconjunto S debe ser consistente.
2. Todas las instancias tiene que ser más cercanas a un vecino selectivo de la misma clase que a cualquier otra instancia de otra clase.
3. No puede haber ningún subconjunto S' que satisfaga las condiciones 1 y 2, y que al mismo tiempo contenga un menor número de instancias que el Subconjunto Selectivo, SS.

La segunda condición es la diferencia principal entre el algoritmo de Hart y el subconjunto calculado con el algoritmo del subconjunto selectivo.

De forma que se puede reformular el punto número dos de la siguiente manera:

Todas las instancias del conjunto de entrenamiento T deben de ser más cercanos a un vecino condensado — miembro del subconjunto condensado $\mathcal{CS}$ — de la misma clase que a cualquier otra instancia de otra clase diferente del $\mathcal{CS}$.

Pudiendo verse como una reformulación del punto nº 1. Además, el punto nº 2 para el subconjunto selectivo permite un subconjunto de menor tamaño, eliminando la necesidad de calcular todas las permutaciones de las instancias en T. Con unos criterios más específicos que los que se encuentran en el algoritmo condensado, el subconjunto selectivo resultante no tiene porqué ser mínimamente consistente. Asimismo, de forma general, no será un subconjunto reducido del S producido por *CNN*.

Algoritmo Subconjunto Selectivo Modificado

Barandela [2] 2005 propuso el algoritmo de subconjunto selectivo modificado, *MSS*, como su propio nombre indica, se trata de una modificación del algoritmo propuesto previamente, ver 3.4. Debido a que no se puede garantizar que el algoritmo de Ritter [26] devuelva un subconjunto, S, el cual sea mínimo y consistente, habiendo sido categorizado como un problema NP-Completo [31], por lo que *MSS* pretende conseguir el subconjunto mínimo consistente mediante el uso de la propiedad selectiva.

La aproximación realizada por Barandela *et al.* modifica la condición nº 3 anteriormente propuesta, mientras que las nº 1 y 2 no son modificadas. De forma que la definición nº 3 queda formulada de la siguiente manera:

El Subconjunto Selectivo Modificado, MSS, se define como el subconjunto del conjunto de entrenamiento TS, el cual $\forall x_i \in TS$ aquella instancia de Y_i qiue es más cercano a otra clase que a la de x_i, i.e. el más cercano a su enemigo más cercano.

El objetivo principal de esta modificación es reforzar la condición que debe cumplir el subconjunto reducido para maximizar la aproximación a la frontera de decisión. Quedando definido el algoritmo, ver algoritmo 9, como una alternativa eficiente al algoritmo propuesto por Ritter *et al.*, siendo capaz de seleccionar mejores instancias (más cercanas a la frontera de decisión).

El criterio que sigue *MSS* para determinar la frontera de decisión es la distancia al enemigo más cercano. Con esta medida se puede definir el mejor subconjunto selectivo coimo aquel que contiene el mejor vecino relacionado para cada instancia en el TS. (Mejor $\iff$ Menor distancia a su enemigo más cercano).

Algorithm 9 Algoritmo *Modified Selective Subset*, *MSS*.

Require: Conjunto de entrenamiento X
Ensure: Conjunto editado $S \subset X$

```
1: procedure MSS(X)
2:     S ← ∅
3:     Ordenar las instancias {x_i}_{i=1}^n en función de D_i a su enemigo más
   cercano
4:     for i = 1 hasta n do
5:         add ← false
6:         for j = i hasta n do
7:             if x_j ∈ T ⋀ d(x_i, x_j) < D_j then
8:                 Eliminar x_j de T
9:                 add ← true
10:            end if
11:        end for
12:        if add then
13:            Añadir x_i a S
14:        end if
15:        if T = ∅ then
16:            return S
17:        end if
18:    end for
19: end procedure
```

3.5. Función distancia entre instancias

Una función distancia proporciona la proximidad entre dos instancias en función de todos sus parámetros. Si la distancia que separa dos instancias es cero, ambas instancias son idénticas. Se tiende a trabajar con conjuntos de datos normalizados, i.e. todos los datos son ajustados a una escala común, independientemente de la escala en la que hayan sido medidos, para evitar que atributos/características con mucha varianza puedan «despistar» a los algoritmos.

Existen multitud de métricas para calcular la distancia, pudiendo variar la distancia calculada en función de cuál se aplique. Se van a comentar las más representativas.

- **Distancia de Minkowski.** La distancia de Minkowski es una métrica en el espacio vectorial normalizado. Es una métrica que se puede modificar con facilidad para calcular la distancia entre dos instancias de diferentes maneras.

  1. $p = 1$, cálculo de la distancia de Manhattan.
  2. $p = 2$, cálculo de la distancia Euclídea.
  3. $p = \infty$, cálculo de la distancia de Chebyshov.

  Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- **Distancia de Manhattan** o distancia del taxista. Es una métrica en un espacio vectorial normalizado, calculándose como la suma de los n segmentos verticales u horizontales que unen dos puntos.

  Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

- **Distancia euclidiana** o norma L2 o distancia L2. Es la distancia en línea recta entre dos puntos de datos en el espacio euclidiano.

  Su fórmula normalizada es la siguiente:

$$\mathbb{D}(x, y) = \sqrt{\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{\sigma_i^2}}$$

- **Distancia de Chebyshov** o distancia del tablero de ajedrez. La distancia entre dos puntos es la mayor de sus diferencias a lo largo de cualquiera de sus dimensiones coordenadas.

  Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \max_i (|x_i - y_i|)$$

- **Distancia del Coseno.** Mide la similitud o distancia entre dos vectores calculando el coseno del ángulo que forman.

  Su fórmula es la siguiente:

$$\mathbb{D}(x, y) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}$$

Técnicas y herramientas

Esta parte de la memoria tiene como objetivo presentar las técnicas metodológicas y las herramientas de desarrollo que se han utilizado para llevar a cabo el proyecto. Si se han estudiado diferentes alternativas de metodologías, herramientas, bibliotecas se puede hacer un resumen de los aspectos más destacados de cada alternativa, incluyendo comparativas entre las distintas opciones y una justificación de las elecciones realizadas. No se pretende que este apartado se convierta en un capítulo de un libro dedicado a cada una de las alternativas, sino comentar los aspectos más destacados de cada opción, con un repaso somero a los fundamentos esenciales y referencias bibliográficas para que el lector pueda ampliar su conocimiento sobre el tema.

Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros3, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

[1] Sunith Shetty . Automl for building simple to complex ml pipelines, Sep 2018.

[2] Ricardo Barandela, Francesc J Ferri, and J Salvador Sánchez. Decision boundary preserving prototype selection for nearest neighbor classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(06):787–806, 2005.

[3] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1:2012, 2012.

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[5] Henry Brighton and Chris Mellish. Advances in instance selection for instance-based learning algorithms. *Data mining and knowledge discovery*, 6(2):153–172, 2002.

[6] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. Crisp-dm 1.0: Step-by-step data mining guide. 2000.

[7] Potomac Two Crows Corporation. *Introduction to data mining and knowledge discovery*. Two Crows, 1999.

[8] Sanjoy Dasgupta, Michael L Littman, and David McAllester. Pac generalization bounds for co-training. *Advances in neural information processing systems*, 1:375–382, 2002.

[9] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[10] IBM Cloud Education. What is data mining?, 2021.

[11] Jesper Engelen and Holger Hoos. A survey on semi-supervised learning. *Machine Learning*, 109, 02 2020.

[12] Geoffrey Gates. The reduced nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 18(3):431–433, 1972.

[13] Zoubin Ghahramani and Michael I Jordan. Supervised learning from incomplete data via an em approach. In *Advances in neural information processing systems*, pages 120–127, 1994.

[14] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences.ºn the Move to Meaningful Internet Systems* , pages 986–996. Springer, 2003.

[15] David J Hand. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.

[16] Peter Hart. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14(3):515–516, 1968.

[17] Mathworks Inc. Supervised learning.

[18] JavaTPoint. Introduction to semi-supervised learning - javatpoint.

[19] Vijay Kotu and Bala Deshpande. Chapter 2 - data mining process. In Vijay Kotu and Bala Deshpande, editors, *Predictive Analytics and Data Mining*, pages 17–36. Morgan Kaufmann, Boston, 2015.

[20] Erik G Learned-Miller. Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts*, 2014.

[21] Cen Li and Gautam Biswas. Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):673–690, 2002.

[22] Onesmus Mbaabu. Clustering in unsupervised machine learning.

[23] J Arturo Olvera-López, J Ariel Carrasco-Ochoa, J Francisco Martínez-Trinidad, and Josef Kittler. A review of instance selection methods. *Artificial Intelligence Review*, 34(2):133–143, 2010.

[24] Alfonso Palmer, Rafael Jiménez, and Elena Gervilla. Data mining: Machine learning and statistical techniques. *Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.)*, pages 373–396, 2011.

[25] Joel Ratsaby and Santosh S Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 412–417, 1995.

[26] G Ritter, H Woodruff, S Lowry, and T Isenhour. An algorithm for a selective nearest neighbor decision rule (corresp.). *IEEE Transactions on Information Theory*, 21(6):665–669, 1975.

[27] Jose Antonio Sanchez. ¿cómo aprenden las máquinas? machine learning y sus diferentes tipos, Aug 2020.

[28] Technovert. Introduction to machine learning, 2020.

[29] Kiran K Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. Attention-based graph neural network for semi-supervised learning. *arXiv preprint arXiv:1803.03735*, 2018.

[30] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284, 2015.

[31] Gordon Wilfong. Nearest neighbor problems. *International Journal of Computational Geometry & Applications*, 2(04):383–416, 1992.

[32] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421, 1972.

[33] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[34] Xueyuan Zhou and Mikhail Belkin. Semi-supervised learning. In *Academic Press Library in Signal Processing*, volume 1, pages 1239–1269. Elsevier, 2014.

[35] Yan Zhou and Sally Goldman. Democratic co-learning. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 594–602. IEEE, 2004.

[36] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.