UNIVERSIDAD DE MURCIA

FACULTAD DE MATEMÁTICAS

Fundamentos Matemáticos del Aprendizaje
Semi-Supervisado

D. Javier Patricio Luján Romero

Mayo 2026

A mi familia por su amor incondicional desde el inicio.

Mi madre y su ternura.

Mi padre y su sacrificio.

Índice general

Abstract III

Resumen V

1. Introducción 1

1.1. Introducción al aprendizaje semi-supervisado 1

1.2. Problema de clasificación semi-supervisado 2

1.3. Suposiciones del aprendizaje semi-supervisado 3

2. Taxonomía de métodos de aprendizaje semi-supervisado 5

2.1. Características discriminantes 5

2.1.1. Suposición sobre la relación entre los datos etiquetados y no etiquetados 5

2.1.2. Inductivo vs transductivo 6

2.1.3. Generativo vs discriminante 6

2.1.4. Basado en el uso de los datos no etiquetados 6

2.2. Taxonomía 7

Bibliografía 11

Abstract

El artículo 5.2 de la *Normativa propia de la Facultad de Matemáticas de la Universidad de Murcia para el Trabajo de Fin de Grado del Grado en Matemáticas* dice:

> La memoria del TFG estará escrita en español, salvo un resumen inicial de al menos 500 palabras en inglés, cuya versión en español se incluirá a continuación.
>
> Para las memorias sujetas a evaluación por tribunal, a petición razonada del/la estudiante y con el visto bueno del tutor/a, la Comisión del TFG podrá autorizar que el resumen esté escrito en otro idioma, o que lo esté la memoria completa, en cuyo caso deberá incluirse un resumen de al menos 500 palabras en ese idioma con su versión en español.

Aquí se incluiría el resumen en inglés.

Resumen

Aquí se incluiría la traducción al español del resumen anterior.

CAPÍTULO 1

Introducción

En este capítulo se introducen los conceptos básicos necesarios para el desarrollo de este trabajo. En la sección 1.1 se realiza una introducción informal que abarca desde el Machine Learning hasta llegar al aprendizaje semisupervisado. Posteriormente, en la sección 1.2 se presenta una definición formal de la clasificación semi-supervisada, el concepto sobre el que se articula el trabajo. Finalmente, en la sección 1.3, se presentan las suposiciones que hay detrás del aprendizaje semi-supervisado, y que son necesarias para entender cómo este puede obtener mejor rendimiento que el aprendizaje supervisado, y de cuándo es así.

1.1. Introducción al aprendizaje semi-supervisado

El **aprendizaje computacional**, o Machine Learning, tiene muchas definiciones, pero de manera intuitiva se puede decir que es el estudio de los métodos computacionales – algoritmos – que mejoran su rendimiento o realizan predicciones haciendo uso de experiencia previa [4].

Esta experiencia previa viene normalmente dada en forma de ejemplos, instancias de los datos que usa el modelo o algoritmo, que normalmente se representan como un vector de características o atributos. Además, es común que cada ejemplo venga acompañado de una etiqueta (*label*), es decir, de un valor o categoría asociado al ejemplo. El conjunto de todos los ejemplos usados por un algoritmo para mejorar su rendimiento se conoce como el conjunto de entrenamiento. [4]

Dentro del aprendizaje computacional, se pueden distinguir dos grandes categorías:

Por un lado, tenemos el **Aprendizaje Supervisado**: Aplicaciones en las que los datos de entrenamiento consisten en ejemplos de vectores de entrada con sus correspondientes etiquetas de salida [1]. En el que a su vez se presenta otra gran subdivisión: “Casos [...] donde el objetivo es asignar cada vector de entrada a un número finito de categorías discretas, son llamados problemas de clasificación. Mientras que, si la salida deseada consiste en una o más variables continuas, entonces la tarea se llama regresión” [1].

Por otro lado, tenemos el **Aprendizaje No Supervisado** cuando “el conjunto de datos de entrenamiento consisten en ejemplos de vectores de entrada sin sus correspondientes etiquetas de salida” [1]. En este caso, el objetivo puede ser encontrar grupos de ejemplos similares en los datos, *clustering*; determinar la distribución de los datos en el espacio de entrada, *density estimation*; o proyectar los datos de un espacio de alta dimensión a uno más sencillo, *dimensionality reduction* [1].

Aparte de estas dos categorías, cabe mencionar otra de gran importancia, el **Aprendizaje por Refuerzo**. Este es bastante diferente a los anteriores, pues se basa en la idea de que el algoritmo de aprendizaje interactúa activamente con el entorno y en algunos casos afecta al entorno, y recibe una recompensa inmediata por cada acción. El objetivo del algoritmo de aprendizaje es maximizar su recompensa a lo largo de un curso de acciones e interacciones con el entorno [4]. Aunque no vayamos a profundizar en el mismo en este trabajo, es importante mencionarlo al ser una rama importante del aprendizaje computacional, y también es importante mencionar su principal dilema, el *exploración vs explotación*: la búsqueda del equilibrio entre explorar nuevas acciones para obtener más información, y explotar las acciones que ya conoce para maximizar su recompensa [1].

Sin embargo, en este trabajo nos vamos a centrar en una rama del aprendizaje computacional distinta, y que se encuentra a medio camino entre los tipos de aprendizaje principales. El **Aprendizaje Semi-Supervisado**, la rama del aprendizaje computacional que usa tanto datos etiquetados como no etiquetados para realizar distintas tareas de aprendizaje [7].

Aun así, esta definición de aprendizaje semi-supervisado es demasiado general para los objetivos de este trabajo, es por esto que se va a tomar el enfoque propuesto por Seeger (2006) en "A taxonomy for semi-supervised learning methods"[6], donde se considera al aprendizaje semi-supervisado como una extensión del aprendizaje supervisado que se beneficia de la presencia de datos no etiquetados. Denominando como **semi-unsupervised learning** a aquellas tareas de aprendizaje no supervisado que se benefician de la presencia de datos etiquetados.

Este tipo de aprendizaje es muy atractivo por dos razones: por un lado, porque puede potencialmente utilizar tanto datos etiquetados como no etiquetados para lograr un mejor rendimiento que el aprendizaje supervisado. Y, por otro lado, porque puede lograr el mismo nivel de rendimiento que el aprendizaje supervisado, pero con menos instancias etiquetadas [9]. Lo cual es muy interesante debido al coste que supone etiquetar los datos –anotadores humanos expertos en la materia en específico–, frente a la abundancia y el bajo coste de obtención de los datos no etiquetados.

Finalmente, dentro del aprendizaje semi-supervisado, al ser una extensión del aprendizaje supervisado, se vuelve a encontrar la subdivisión entre problemas de clasificación y regresión. Este trabajo se va a centrar únicamente en el aprendizaje semi-supervisado para problemas de clasificación, aunque muchos de los resultados que se presentan también se pueden aplicar al aprendizaje semi-supervisado para problemas de regresión.

1.2. Problema de clasificación semi-supervisado

Una vez introducido el concepto principal del trabajo, el aprendizaje semi-supervisado, y dado que este trabajo se va a centrar en el aprendizaje semi-supervisado para problemas de clasificación, es necesario definir de manera formal este problema. Para ello, vamos a hacer uso de una serie de definiciones que se encuentran en el libro de Xiaojin Zhu, "Introduction to Semi-Supervised Learning" [9], y que se presentan a continuación.

Como primer paso se presenta la definición del aprendizaje supervisado, que, como se ha mencionado, es la base del aprendizaje semi-supervisado:

Definición 1.2.1 (Aprendizaje Supervisado) Sea $\mathscr{X}$ el dominio de las instancias, e $\mathscr{Y}$ el dominio de las etiquetas. Sea $P(\mathbf{x}, y)$ una distribución de probabilidad conjunta (desconocida)

sobre las instancias y etiquetas $\mathscr{X} \times \mathscr{Y}$. Dado un conjunto de entrenamiento $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ con $(\mathbf{x}_i, y_i) \overset{\text{i.i.d.}}{\sim} P(\mathbf{x}, y)$ (independiente e idénticamente distribuido), el aprendizaje supervisado entrena una función $f : \mathscr{X} \to \mathscr{Y}$ en alguna familia de funciones $\mathscr{F}$, con el objetivo de que $f(\mathbf{x})$ prediga la verdadera etiqueta y en datos futuros $\mathbf{x}$, donde $(\mathbf{x}, y) \overset{\text{i.i.d.}}{\sim} P(\mathbf{x}, y)$ también.

Partiendo de esta definición de aprendizaje supervisado, se puede entonces definir el problema de clasificación:

Definición 1.2.2 (Clasificación) La clasificación es el problema de aprendizaje supervisado con clases discretas $\mathscr{Y}$. La función f se denomina clasificador.

Para finalmente poder formalizar el problema de clasificación dentro del aprendizaje semi-supervisado:

Definición 1.2.3 (Clasificación Semi-Supervisada) Es una extensión del problema de clasificación supervisada. Los datos de entrenamiento consisten tanto en l instancias etiquetadas $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ como en u instancias no etiquetadas $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$. Se asume típicamente que hay muchos más datos no etiquetados que etiquetados, es decir, $u \gg l$. El objetivo de la clasificación semi-supervisada es entrenar un clasificador f a partir de los datos etiquetados y no etiquetados, de tal manera que sea mejor que el clasificador supervisado entrenado únicamente con los datos etiquetados.

De esta manera, queda establecido de manera formal el marco teórico sobre el que se va a desarrollar este trabajo. En los siguientes capítulos se va a, primero, presentar una taxonomía para los métodos de clasificación semi-supervisada, para posteriormente desarrollar algunos de estos con mayor profundidad. Pero antes de terminar este capítulo, se va a presentar una idea de cómo el aprendizaje semi-supervisado obtiene mejor rendimiento que el aprendizaje supervisado, y de cuándo es así.

1.3. Suposiciones del aprendizaje semi-supervisado

Aunque no todos los métodos de aprendizaje semi-supervisado se basan explícitamente en probabilidades, resulta útil para entender cómo el aprendizaje semi-supervisado puede obtener mejor rendimiento que el aprendizaje supervisado, asumir que los métodos representan las hipótesis como $p(y \mid x)$ y los datos no etiquetados como $p(x)$. De esta manera, la información que proporciona $p(x)$ se puede introducir de dos posibles formas: o bien en modelos que estiman la distribución conjunta $p(x, y)$ que comparte parámetros con $p(x)$, o bien modificando la función objetivo para incluir términos de $p(x)$ [10]. En ambos casos, el aprendizaje semi-supervisado puede obtener mejor rendimiento que el aprendizaje supervisado, porque puede aprovechar la información contenida en los datos no etiquetados para mejorar la estimación de $p(y \mid x)$.

Sin embargo, aunque a priori parece que el aprendizaje semi-supervisado siempre debería obtener mejor rendimiento que el aprendizaje supervisado, esto no es así. Detrás de cada método de aprendizaje semi-supervisado hay una serie de suposiciones sobre la relación entre $p(x)$ y $p(y \mid x)$, y si estas suposiciones no se cumplen, el aprendizaje semi-supervisado puede obtener un rendimiento peor que el aprendizaje supervisado [9].

Veamos entonces, de manera breve, cuáles son estas suposiciones como las presenta Chappel et al. (2006) en "Semi-supervised learning" [3]:

- **Suposición de suavidad**: Si dos puntos x_1 y x_2 en una región de alta densidad están cerca, entonces también deben estarlo sus correspondientes salidas y_1 y y_2.
- **Suposición de clústers**: Si los puntos se encuentran en un mismo clúster, es probable que tengan la misma etiqueta.
- **Suposición de baja densidad**: La frontera de decisión debe pasar por regiones de baja densidad. Esta suposición es una reformulación de la suposición de clústers pues si los puntos de un mismo clúster tienen la misma etiqueta, entonces la frontera de decisión debe pasar por regiones de baja densidad donde no haya clústers.
- **Suposición de variedad subyacente**: Los datos, de alta dimensión, se encuentran (aproximadamente) en un manifold de baja dimensión.

Conforme se vayan presentando los distintos métodos de clasificación semi-supervisada, se irá viendo cuáles de ellos se basan en una o varias de estas suposiciones, y cómo el rendimiento de estos método dependerá de si estas suposiciones se cumplen o no en los datos con los que se trabaja.

CAPÍTULO 2 Taxonomía de métodos de aprendizaje semi-supervisado

En este capítulo se va a presentar una taxonomía de los diferentes métodos de clasificación bajo el enfoque de aprendizaje semi-supervisado. Su objetivo es permitir un entendimiento claro de la base de los diferentes métodos con el fin de facilitar su comprensión y estudio. Para realizarla se ha utilizado como estructura principal la taxonomía presentada en el artículo *A survey on Semi-supervised Learning, van Engelen y Hoos, 2020* [7] a la que se le han realizado ciertas modificaciones debidamente motivadas buscando una mejor comprensión y adaptación a los objetivos de este trabajo.

2.1. Características discriminantes

Antes de exponer la taxonomía, se van a introducir posibles características discriminantes que pueden ser utilizadas para clasificar los diferentes métodos de aprendizaje semi-supervisado, explicando para cada una de ellas su significado y su importancia a la hora de clasificar los diferentes métodos. Finalmente, la taxonomía se creará a partir de estas características, asignando una jerarquía de niveles a las mismas y agrupando los diferentes métodos en las hojas del árbol resultante.

2.1.1. Suposición sobre la relación entre los datos etiquetados y no etiquetados

Como se explicó al final del último capítulo las diferentes suposiciones sobre cómo se relacionan los datos etiquetados y no etiquetados es una característica básica a la hora de entender los diferentes métodos de aprendizaje semi-supervisado. Es por esto, que surge rápidamente como una posible característica discriminante a la hora de clasificar los diferentes métodos, como por ejemplo se hace en el libro de Chapelle *Semi-supervised learning* [3].

Sin embargo, clasificar en base a esta característica no es ni tan sencillo ni tan claro como parece a primera vista. Por un lado, las suposiciones teóricas a menudo se solapan conceptualmente, o son compartidas por diferentes algoritmos cuya implementación matemática difiere radicalmente. Además, hay ciertos algoritmos de aprendizaje semi-supervisado – como los métodos *wrapper* que se presentarán más adelante–, que no se basan en suposiciones explícitas sobre los datos, por lo que una clasificación por esta característica resulta incompleta. Por tanto, aunque es importante conocer la suposición de cada algoritmo a la hora de trabajar con él, una clasificación en base a esta característica puede no resultar suficientemente esclarecedora para entender las ideas y fundamentos en los que se basa.

2.1.2. Inductivo vs transductivo

Una segunda característica discriminante, que es, de hecho, un estándar de facto en la literatura a la hora de clasificar los diferentes métodos de aprendizaje semi-supervisado, es la distinción entre métodos inductivos y transductivos. Esta distinción se basa en el objetivo final del método: Por un lado, los **métodos inductivos** buscan aprender una función de predicción $f : \mathscr{X} \to \mathscr{Y}$ que pueda ser aplicada a cualquier punto del espacio de entrada $\mathscr{X}$, mientras que los **métodos transductivos** se centran en realizar predicciones únicamente para un conjunto específico de puntos de *prueba*, los puntos no etiquetados de la entrada [3].

Esta distinción es crucial pues refleja el objetivo final de un método, y por tanto influye en su diseño y en las técnicas utilizadas para su implementación. Mientras que los métodos trasductivos pueden aprovechar al máximo la información de los datos no etiquetados para mejorar sus predicciones, los métodos inductivos deben ser capaces de generalizar a nuevos puntos de entrada, lo que puede requerir técnicas adicionales para evitar el sobreajuste a los datos de entrenamiento.

2.1.3. Generativo vs discriminante

Otro criterio de clasificación estándar en la literatura, sobre todo la más clásica, es la distinción entre métodos generativos y discriminantes. Esta distinción se presentó de manera implícita en el capítulo anterior, cuando se explicó cómo la relación entre $p(x)$ y $p(y|x)$ es la razón por la que el aprendizaje semi-supervisado mejora sus predicciones.

Este criterio clasifica los métodos en dos tipos: Los métodos **generativos** que buscan aprender un modelo de la probabilidad conjunta de los datos $p(x,y)$, de las entradas x, y de las etiquetas y, para hacer predicciones usando la regla de Bayes para calcular $p(y|x)$ y seleccionando la etiqueta más probable. Y los métodos **discriminantes** que se centran en modelar directamente la probabilidad a posteriori $p(y|x)$ o aprenden un mapeo directo de las entradas a las etiquetas [5].

Esta distinción es muy útil pues permite comprender los fundamentos que hay detrás de un método, antes de abordar los detalles de su implementación concreta. Sin embargo, no es una distinción perfecta, ya que no contempla todos los posibles métodos, en particular, los métodos que se basan en la geometría o topología de los datos –como los métodos basados en grafos o variedades–.

2.1.4. Basado en el uso de los datos no etiquetados

Finalmente, se presenta una última característica discerniente, la forma en la que los diferentes algoritmos hacen uso de los datos no etiquetados. Esta característica es la base de la taxonomía que se presenta en el trabajo de van Engelen y Hoos [7], y por ende es la base de la taxonomía presentada en este trabajo.

Esta característica clasifica los algoritmos de aprendizaje semi-supervisado en tres grupos:

- Los métodos **Envolventes** o **Wrapper**, métodos que se basan en un paso de pseudo-etiquetado. Es decir, se basan en un proceso que utiliza uno o varios algoritmos supervisados, que parten solo con los datos etiquetados, y va iterativamente etiquetando parte de

los datos no etiquetados usando estos algoritmos, para posteriormente reentrenarlos usando estas nuevas etiquetas, sin que estos distingan la diferencia entre etiquetas iniciales y etiquetadas por el proceso.

- Los métodos basados en **preprocesado no supervisado**, aquellos que se basan en un pipeline de dos etapas: una primera etapa de preprocesamiento no supervisado, que puede ser por ejemplo una etapa de reducción de dimensionalidad o de extracción de características, seguida de una etapa de clasificación donde un algoritmo supervisado, agnóstico a la etapa anterior, se entrena con los datos etiquetados y la representación obtenida en la primera etapa.
- Los métodos **intrínsecamente semi-supervisados**, que directamente optimizan una función objetivo haciendo uso tanto de los datos etiquetados como los no etiquetados. Por tanto, no necesitan pasos intermedios ni algoritmos supervisados de base, sino que usualmente son extensiones de estos algoritmos supervisados que integran de manera explícita los datos no etiquetados en su formulación matemática.

El empleo de esta característica para clasificar es de gran utilidad, ya que permite realizar una división funcional de los métodos clara y sencilla, que permite entender la estructura general detrás de un método, sin necesidad de abordar los fundamentos matemáticos del mismo.

2.2. Taxonomía

Basándose en todo lo explicado anteriormente, se presenta en la figura 2.1 una taxonomía de, principalmente, dos niveles de profundidad, que profundiza a un tercer nivel en los métodos inductivos intrínsecamente semi-supervisados.

1. Primero, se distingue entre los métodos inductivos o transductivos.
2. Una vez clasificados en base a esta característica, se considera cómo el algoritmo hace uso de los datos no etiquetados.
3. Finalmente, dentro de los métodos inductivos e intrínsecamente semi-supervisados, se realiza una última distinción basada en el fundamento matemático del método, distinguiendo entre métodos generativos y discriminantes, añadiendo además una tercera rama para los métodos geométricos (basados en variedades).

Cabe notar, como se viene comentando, la similitud entre esta taxonomía y la presentada en el trabajo de van Engelen y Hoos [7]. Se va primero a explicar las similitudes, que son las que forman la base de la taxonomía presentada, para posteriormente explicar las diferencias que hay entre ambas.

La primera similitud es la división inicial de la taxonomía en métodos inductivos y transductivos, esta división, estándar de facto en la literatura, es, como se ha comentado en la sección 2.1.2, crucial para entender el objetivo final de un método y por tanto es lo primero que se ha de tener en cuenta a la hora de abordar un método de aprendizaje semi-supervisado.

La segunda similitud, y la base de la taxonomía de van Engelen y Hoos [7], es la división de los métodos inductivos en base a cómo hacen uso de los datos no etiquetados, esta división, como se ha comentado en la sección 2.1.4, permite entender la estructura general detrás de un método, siendo una herramienta muy útil para comprender de antemano un método o algoritmo.

Estas similitudes fundamentan la base de una taxonomía sólida y clara, que permite entender como va actuar de manera general un método, sin necesidad de abordar los detalles matemáticos del mismo.

Sin embargo, y aquí aparece la primera diferencia, para realizar esta taxonomía, se ha decidido seguir otro estándar de la literatura, y distinguir entre los métodos discriminantes y generativos, como se ha explicado en la sección 2.1.3. Esta distinción permite entender mejor los fundamentos matemáticos sobre los que se basa cada método, y aunque no engloba todos los métodos, al realizarse únicamente dentro de los métodos inductivos intrínsecamente semi-supervisados, no resulta tan limitante como si se realizara a un nivel más alto de la taxonomía.

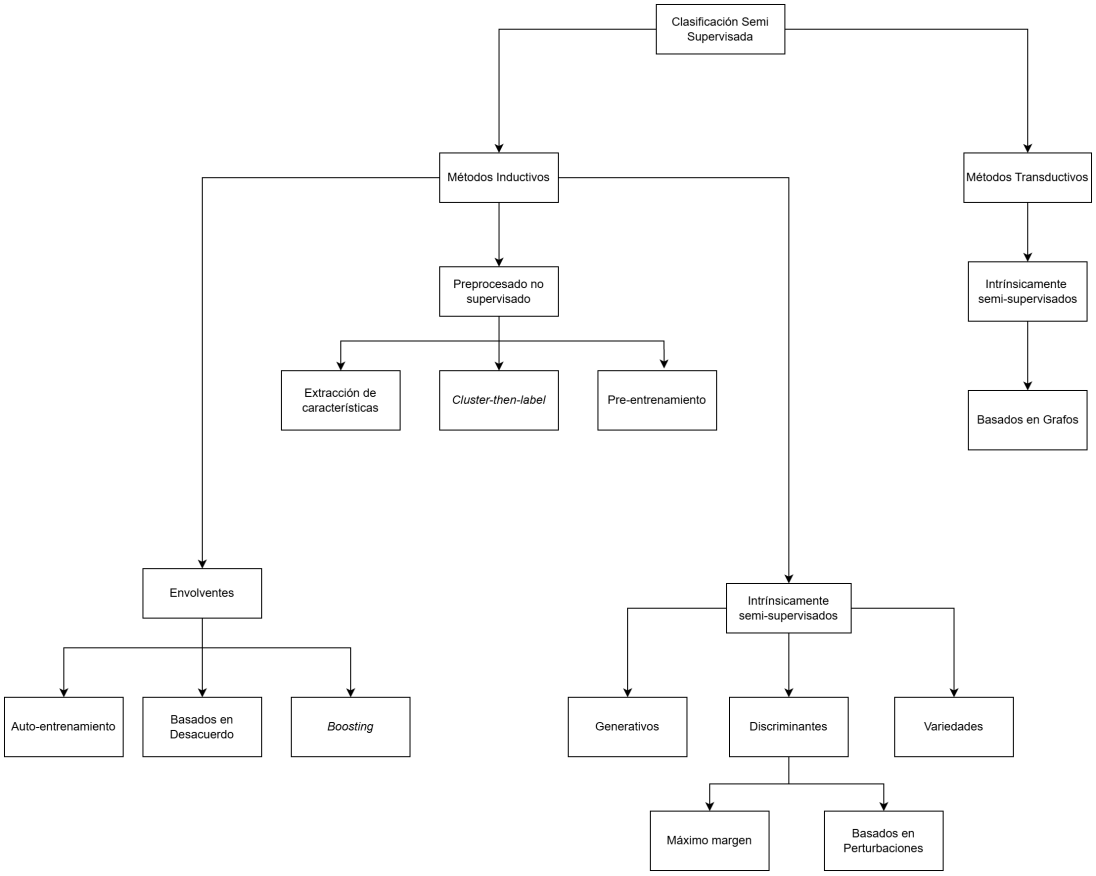

Figura 2.1: Taxonomía de métodos de aprendizaje semi-supervisado.

De esta forma, la taxonomía propuesta reorganiza los métodos intrínsecos en tres ramas diferenciadas: la familia de métodos generativos; la de métodos discriminantes —que agrupa bajo un mismo paraguas conceptual a los enfoques de máximo margen y a los basados en perturbaciones—; y finalmente, una categoría independiente para los métodos basados en variedades, reconociendo así su naturaleza geométrica distintiva.

Además de esta diferencia, se han añadido dos cambios menores:

Por un lado, para mantener la simetría entre los métodos inductivos y transductivos, se ha añadido en la rama de los métodos transductivos la división sobre el uso de los datos no etiquetados, aunque en este caso solo hay algoritmos intrínsecamente semi-supervisados. De esta manera se mantiene la simetría entre ambas ramas, y permite tener una mayor comprensión con tan solo ver la figura.

Por último, dentro de los métodos envolventes, se ha cambiado el nombre de la familia presentada por van Engelen y Hoos [7] como co-training por *métodos basados en desacuerdo*. De esta manera, se evita nombrar a la familia por un algoritmo en concreto dentro de esta, el algoritmo de co-training de Blum y Mitchell [2]. El objetivo del nuevo nombre, que se desarrollará en profundidad en el siguiente capítulo, es reflejar la idea general de esta familia de métodos tal y como se presenta en el trabajo de Zhou y Li [8].

Bibliografía

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[2] 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.

[3] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 09 2006.

[4] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[5] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

[6] Matthias W Seeger et al. A taxonomy for semi-supervised learning methods., 2006.

[7] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

[8] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

[9] Xiaojin Zhu and Andrew Goldberg. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.

[10] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.