Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Bosque Aleatorio (*Random Forest*)

1

Profesora: *Dra. Fabiola Ocampo Botello*

Evaluación de los modelos creados

Una vez que se ha creado un modelo de clasificación con datos previos, es importante tener presente cómo se comportará este clasificador con datos futuros.

Esto significa tener una estimación de la precisión de su comportamiento en el futuro, con datos con los cuales el clasificador no ha sido entrenado, no conoce.

2

Rokach, L. & Maimon, O. (2015) establecen que la meta de un algoritmo de clasificación se puede definir formalmente como:

Dado un conjunto de entrenamiento S con atributos de entrada

$$A = \{a1, a2, \ldots, an\}$$

y un atributo nominal *y* y una distribución desconocida D, la meta es inducir un clasificador óptimo con el mínimo error de generalización.

3

4

Estimación empírica del error de generalización

Uno de los enfoques para estimar el error de generalización es el método de retención (*holdout method*) en el que el conjunto de datos dado se divide aleatoriamente en dos conjuntos: Conjuntos de entrenamiento y prueba (Rokach, L. & Maimon, O, 2015) .

Por lo general, dos tercios de los datos se consideran para el conjunto de entrenamiento y los datos restantes se asignan al conjunto de prueba.

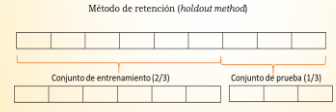

Imagen: Elaboración propia

El submuestreo aleatorio (*Random subsampling*) y la validación cruzada n-fold (*n-fold cross-validation*) son dos métodos comunes de remuestreo (Rokach, L. & Maimon, O, 2015):

- En el submuestreo aleatorio, los datos se dividen aleatoriamente varias veces en conjuntos de entrenamiento y pruebas disjuntos. Los errores obtenidos de cada partición se promedian.
- En la validación cruzada n-fold, los datos se dividen aleatoriamente en n subconjuntos mutuamente excluyentes de aproximadamente el mismo tamaño. Un inductor es entrenado y probado n veces; cada vez se prueba en uno de los k pliegues (fold) y se entrena utilizando los n-1 pliegues (fold) restantes.

5

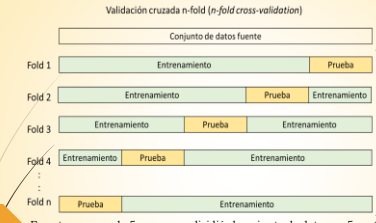

Imagen: Elaboración propia

En este caso n vale 5, ya que se dividió el conjunto de datos en 5 partes (fold1, fold2, fold3, fold4, fold5), por citar:
Modelo 1: entrenado en fold1+fold2+fold3+fold4 y probado en fold5
Modelo 2: entrenado en fold1+fold2+fold3+fold5 y probado en fold4
Modelo 3: entrenado en fold1+fold2+fold4+fold5 y probado en fold3
Modelo 4: entrenado en fold1+fold3+fold4+fold5 y probado en fold2
Modelo 5: entrenado en fold2+fold3+fold4+fold5 y probado en fold1

6

Evaluación mediante validación cruzada

Imagen tomada de Hernández, Ramírez y Ferri (2004).

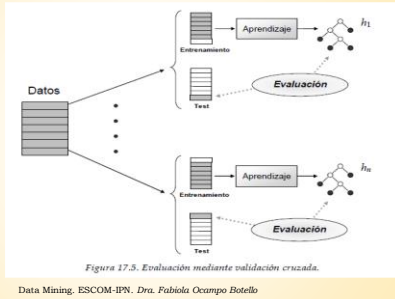

Figura 17.5. Evaluación mediante validación cruzada.

7

Evaluación por bootstrap

Es un método similar al de validación cruzada.

Suponga que se tienen N ejemplos, del cual se realiza un muestreo aleatorio con reposición, esta muestra será el conjunto de entrenamiento, debido a que el muestreo se realizó con reemplazo, puede contener elementos repetidos, los cuales se mantienen, significa también que no contendrá algunos elementos del conjunto original. *Los elementos no elegidos se utilizan para realizar la prueba*. El proceso anterior se repite un número prefijado k de veces (digamos diez veces) y después se actúa como en el caso de la validación cruzada, promediando los errores/precisiones (Hernández, Ramírez y Ferri, 2004).

Teóricamente, esto da un conjunto de entrenamiento de N ejemplos y un conjunto de test de aproximadamente 0,368 × N ejemplos (Hernández, Ramírez y Ferri, 2004). El tercio restante se le conoce como *out of bag*.

8

Sesgo y Varianza (Bias & Variance)

Amat (2020) establece que los modelos de aprendizaje estadístico y los de aprendizaje automático tienen el problema de el equilibrio entre sesgo y varianza.

Sesgo	Varianza
- Se refiere a qué tanto se alejan en promedio las predicciones de un modelo respecto a los valores reales. - Refleja qué tan capaz es el modelo de aprender la relación verdadera, real que existe entre los predictores y la variable de respuesta. - Por ejemplo, si se tiene una distribución con distribución no lineal y se utiliza un modelo de regresión lineal, habrá mucho sesgo.	- Se refiere a cuánto cambia el modelo dependiendo de los datos utilizados en su entrenamiento. - Puede ser que el modelo memorice los datos en lugar de aprender la verdadera relación entre los predictores y la variable respuesta. - Por ejemplo, un modelo de árbol con muchos nodos, suele variar su estructura con que apenas cambien unos pocos datos de entrenamiento, tiene mucha varianza

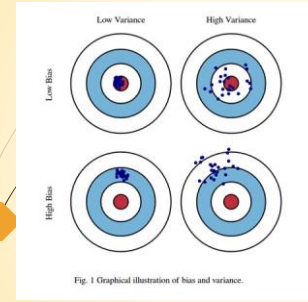

Fig. 1 Graphical illustration of bias and variance.

Imagen Creative Commons
En:
https://un1crom.medium.com/reconsider-time-3686308a5e69

Sesgo (Bias) qué tanto se alejan en promedio las predicciones de un modelo respecto a los valores reales.

Varianza (Variance) qué tanto se dispersan (en promedio) los datos de su media.

¿Cómo se controlan el sesgo y varianza en los modelos basados en árboles?

Amat (2020) menciona que:

- Por lo general, los ***árboles pequeños*** (pocas ramificaciones) tienen poca varianza pero no consiguen representar bien la relación entre las variables, es decir, tienen sesgo alto.
- En contraposición, los ***árboles grandes*** se ajustan mucho a los datos de entrenamiento, por lo que tienen muy poco sesgo pero mucha varianza. Una forma de solucionar este problema son los métodos de ensemble.

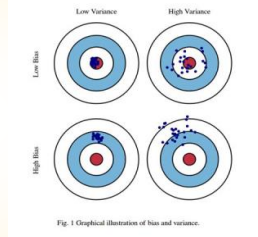

Fig. 1 Graphical illustration of bias and variance.

ImageEn: https://un1crom.medium.com/reconsider-time-3686308a5e69
n Creative Commons

¿Cómo aumentar la precisión de un clasificador?

Considerando que uno de los problemas que existen es el desequilibrio de clases.

Mediante los métodos de conjuntos (métodos de ensamble). Un conjunto (ensamble) de clasificación es un modelo compuesto que está formado por una combinación de clasificadores. Los clasificadores individuales votan y el conjunto devuelve una predicción de etiqueta de clase basada en la recopilación de votos (Han, Kamber & Pei, 2012).

El bagging, el boosting y los bosques aleatorios (random forerst) son ejemplos de métodos de conjunto (Han, Kamber & Pei, 2012).

13

Para entender la importancia de un conjunto de clasificación, considere el ejemplo propuesto por Hernández, Ramirez y Ferri (2004):

Suponga que tiene un conjunto formado por tres clasificadores {h1, h2, h3} y sea x un nuevo dato a ser clasificado.
Sí los tres clasificadores son similares, entonces cuando h1(x) sea erróneo, probablemente h2(x) y h3(x) también lo serán. Sin embargo, si los clasificadores son lo bastante diversos, los errores que cometan estarán poco correlacionados, y por tanto, cuando h1(x) sea erróneo, h2(x) y h3(x) podrían ser correctos, y entonces, si la combinación se realizase por votación mayoritaria, el conjunto combinado clasificaría correctamente el dato x.

14

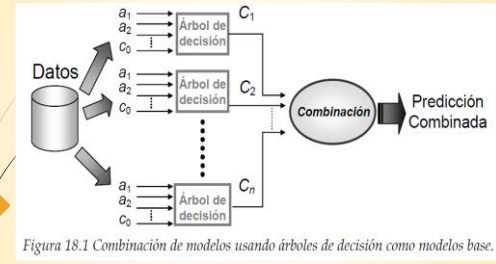

Figura 18.1 Combinación de modelos usando árboles de decisión como modelos base.

Imagen tomada de Hernández, Ramírez y Ferri (2004).

15

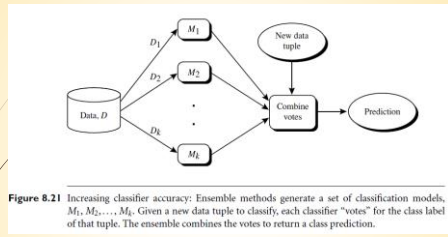

Figure 8.21 Increasing classifier accuracy: Ensemble methods generate a set of classification models, $M_1, M_2, \ldots, M_k$. Given a new data tuple to classify, each classifier "votes" for the class label of that tuple. The ensemble combines the votes to return a class prediction.

Figura tomada de Han, Kamber, & Pei (2012).

16

Bagging

El término *bagging* deriva del mecanismo denominado *bootstrap aggregation*, mecanismo que genera subconjuntos de entrenamiento seleccionando aleatoriamente y con reemplazamiento Dado que hay un conjunto de clasificadores, la predicción de nuevos ejemplos se efectúa por votación mayoritaria. (Hernández, Ramírez y Ferri, 2004).

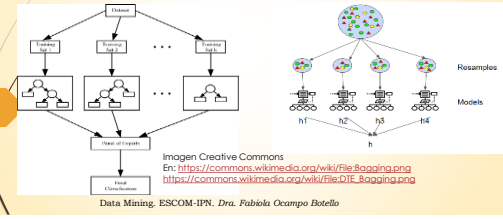

Imagen Creative Commons
En: https://commons.wikimedia.org/wiki/File:Bagging.png
https://commons.wikimedia.org/wiki/File:DTE_Bagging.png

Boosting

Hernández, Ramírez y Ferri (2004) establecen que:

- La estrategia de *boosting* construye los nuevos modelos tratando de corregir los errores cometidos previamente.
- Existen muchas variantes del algoritmo básico de *boosting*, siendo probablemente AdaBoost una de las versiones originales y todavía más populares.
- A diferencia del algoritmo *Bagging*, este algoritmo no siempre realiza las k iteraciones requeridas por el usuario, dado que considera un criterio de parada de acuerdo con el error e.

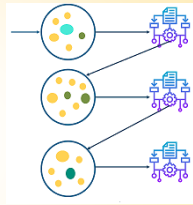

Imagen Creative Commons
En:
https://machinelearningparatodos.com/cual-es-la-diferencia-entre-los-metodos-de-bagging-y-los-de-boosting/

Se aprende iterativamente una serie de k clasificadores. Después de que se aprende un clasificador, Mi, los pesos se actualizan para permitir que el clasificador posterior, Mi+1, "preste más atención" a las tuplas de entrenamiento que Mi clasificó erróneamente (Han, Kamber & Pei, 2012).

Se suman los pesos de cada clasificador que asignó la clase c a X. La clase con la suma más alta es la "ganadora" y se devuelve como la predicción de clase para la tupla X (Han, Kamber & Pei, 2012:382).

Amat (2020) menciona que tres de los métodos de boosting más empleados son AdaBoost, Gradient Boosting y Stochastic Gradient Boosting

Han, Kamber & Pei (2012) establecen que:

- AdaBoost (abreviatura de *Adaptive Boosting*) es un algoritmo de (*boosting*) impulso popular.
- Si una tupla se clasificó incorrectamente, su peso aumenta. Si una tupla se clasificó correctamente, se reduce su peso.
- El peso de una tupla refleja lo difícil que es clasificar: cuanto mayor es el peso, más a menudo se clasifica erróneamente.
- Estos pesos se utilizarán para generar las muestras de entrenamiento para el clasificador de la siguiente ronda.
- La idea básica es que cuando se construya un clasificador, queremos que se centre más en las tuplas mal clasificadas de la ronda anterior.

AdaBoost ha sido definido para problemas de clasificación (aunque se puede adaptar a regresión) a diferencia de bagging. (Hernández, Ramírez y Ferri, 2004).

Ejemplo ilustrativo para entender la diferencia entre bagging y boosting (Han, Kamber & Pei, 2012):

Suponga que es un paciente y le gustaría que le hicieran un diagnóstico basado en sus síntomas.

Bagging	Boosting
En lugar de preguntarle a un médico, puede optar por preguntar a varios. Si se produce un determinado diagnóstico más que cualquier otro, puede elegir este como el diagnóstico final o mejor. Es decir, el diagnóstico final se realiza por mayoría de votos, donde cada médico obtiene un voto igual.	Suponga que asigna ponderaciones al valor o el valor del diagnóstico de cada médico, en función de la precisión de los diagnósticos anteriores que han realizado. El diagnóstico final es entonces una combinación de los diagnósticos ponderados.

Random Forest

Imagine que cada uno de los clasificadores del conjunto es un clasificador de árbol de decisión, de modo que la colección de clasificadores es un "bosque". Los árboles de decisión individuales se generan utilizando una selección aleatoria de atributos en cada nodo para determinar la división (Han, Kamber, & Pei, 2012:383).

Imagen Creative Commons
En: https://www.flickr.com/photos/worldbank/14431641O9

Un conjunto de bosque aleatorio utiliza una gran cantidad de árboles de decisión individuales sin podar que se crean aleatorizando la división en cada nodo del árbol de decisión (Rokach, L. & Maimon, O., 2015).

Aunque el bosque aleatorio se definió para árboles de decisión, este enfoque es aplicable a todos los tipos de clasificadores (Rokach, L. & Maimon, O., 2015:126).

Los bosques aleatorios se pueden construir usando ensacado (bagging) (Han, Kamber, & Pei, 2012:383).

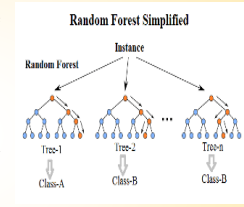

Imagen Creative Commons
En: https://sefiks.com/2017/11/19/how-random-forests-can-keep-you-from-decision-tree/

Referencias bibliográficas

Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). Data Mining: concepts and techniques. Third edition. Morgan Kaufman Series.

Hernández Orallo, José; Ramírez Quintana, Ma. José y Ferri Ram, César. (2004). Introducción a la Minería de Datos. Pearson-Prentice–Hall.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.