Tema 2. : Metodología de la minería de datos.

2.1. Aprendizaje inductivo y modelado
2.2. Entrenamiento y validación (Training y test)
2.3. Validación e interpretación. Procedimientos de validación
2.4. Expresividad: Ajuste, sobreajuste y subajuste

Los distintos métodos o técnicas de la minería de datos no suponen en ningún caso soluciones únicas, ni unívocas, para las distintas tareas o problemas. De hecho, distintas *técnicas* pueden utilizarse para distintas *tareas*, quizás con algunas adaptaciones, y una misma tarea puede ser abordada con distintas técnicas. Este hecho nos plantea algunas cuestiones que acompañan a la metodología de la minería de datos: ¿qué tienen de común las técnicas? ¿qué tienen de específico? ¿hasta qué punto , cada técnica es más o menos valida, según la tarea y la información de cada problema? ¿cómo dan cuenta , "expresan", la información de los datos? ¿qué técnica debemos emplear, en definitiva , en cada caso concreto?

En este tema intentaremos si no dar una contestación definitiva a estas pregunta sí, indicar, al menos, la estrategia , los caminos, para abordarlas. Todo ello, sin perder de vista que la minería de datos, por su propia naturaleza es una disciplina muy empírica y que resuelve sus problemas metodológicos optando por las mejores opciones de entre las disponibles en cada caso concreto. Una visión más completa de los conceptos introducidos en este tema puede verse en (Hernández, Ramirez, M, & Ferri, C, 2010)

2.1.Aprendizaje inductivo y modelado.

La minería de datos, como conjunto de técnicas para resolver diferentes tareas de obtención de conocimiento no explícito a partir de los datos, pretende obtener, extraer, un "**patrón**" de los datos: una regularidad, un resumen no explicito, una hipótesis explicativa, un esquema de asignación, etc. De una u otro forma, casi todas las técnicas de minería de datos (excepto quizá las reglas de asociación/correlación) se centran en la idea de "aprendizaje inductivo". Se trata, en definitiva de "aprender" un patrón general (que se pretende válido en términos generales, más allá de la información de la base de datos) inducido a partir de la información (particular) de los datos (ciertamente los datos son muchos pero *particulares*).

La idea de aprendizaje se relaciona con la mejora con la experiencia (Mitchell, 1977) la capacidad de predicción de observaciones futuras (o explicar observaciones pasadas), la identificación de regularidades (patrones), y , según una perspectiva más teórica y matemática (Solomonoff, 1966),se trataría de eliminar la redundancia, de comprimir la información.

En definitiva: **el aprendizaje inductivo permitiría identificar regularidades que pueden representarse por patrones o modelos, que los resumen (los datos), los definen y pueden predecir futuras observaciones o explicar las pasadas, lo que supone una mejora con la experiencia.**

El aprendizaje (inductivo) puede ser **incremental** o **no** ,según si los datos se presentan poco a poco de forma secuencial o no. En minería de datos lo habitual es que los datos se presenten en un base de datos completa, pero en algunos casos los datos pueden ir cambiando con el tiempo o incrementarse la información disponible y existen técnicas particulares para abordar esta situación (Minería de datos en streaming, minería web, por ejemplo)

También puede darse el caso de tener que vérnoslas con datos interactivos que se modifican con el tiempo, no es lo habitual pero en ocasiones es así (Minería web, o en entornos de software, por ejemplo) . Hay también otros tipos particulares de aprendizaje inductivo que superan las pretensiones del curso(query learning o aprendizaje por consultas, aprendizaje por refuerzo y otros)

En cualquier caso debemos considerar que la extracción de un patrón, el aprendizaje inductivo de un patrón , a partir de los datos, no deja de ser la obtención (la generación) de una **hipótesis** , una hipótesis explicativa, comprensiva o predictiva pero una hipótesis que, como ta,l debe ser validada.

2.2. Entrenamiento y validación (Training y test)

2.2.1. Evaluación y validación.

La validación del patrón obtenido es algo que va más allá de la evaluación de los resultados en la resolución del problema concreto al que nos enfrentamos (clasificación, predicción, agrupación, asociación, etc.) pero también lo incluye, por supuesto.

Por un lado un buen modelo debe ser, además de eficaz para el problema (clasifica bien, agrupa bien, predice bien, asocia bien, etc.) de interés para el usuario, más o menos novedoso, interpretable, sencillo y aplicable. Cuestiones estas que , aunque más difíciles de discernir no deben nunca obviarse.

Con todo, una cuestión central en la validación es *la evaluación de su funcionamiento* (**performance**) tanto en términos absolutos como en relación con distintos métodos y problemas.

En este sentido, además de evaluar los modelos o patrones extraídos por cada método es interesante evaluar y comparar entre sí los distintos modelos posibles. Un método debería ser mejor que otro si genera mejores modelos ; más eficientes en la resolución del problema (o problemas) a los que se enfrenta. Lo curioso es que dado cualquier método de aprendizaje (por bueno que sea) siempre existen problemas concretos que son resueltos mejor por otros métodos. Esta característica (tan fundamental/fundacional de la minería de datos) se conoce como “no free lunch theorem” (No hay almuerzo gratis) (Wolpert & Macready, 1997)

Esto hace que la comparación y el uso de técnicas alternativas sea una constante en la práctica de la minería de datos. Sin embargo, sí es cierto que, en algunos casos, hay algunos métodos que tienden a funcionar mejor que otros en algunos tipos de problemas bajo ciertas condiciones.

También es importante considerar que el análisis global del funcionamiento (performance) más o menos eficiente de un método, a veces , debe ir acompañado de otras matizaciones puesto que puede que , en términos globales tenga un buen funcionamiento con una alta precisión media pero con altibajos en distintos conjuntos de datos o con alta sensibilidad a ciertas variaciones , de orden, por ejemplo.

En cualquier caso, la evaluación de un método es siempre una cuestión fundamental que debe llevarse a cabo adecuadamente, tanto para saber, en un nuestro caso concreto, la calidad de nuestro resultado como para poder compararlo con otras alternativas.

2.2.2. Entrenamiento y validación

En este sentido, no debemos olvidar a la hora de analizar la calidad de los resultados que nuestro propósito es la obtención de un patrón general (de clasificación, por

ejemplo) en este sentido queremos que el clasificador obtenido (o el patrón del tipo que sea) nos sirva adecuadamente no sólo para dar cuenta de los datos con los que ha sido obtenido sino que siga cumpliendo su papel correctamente más allá de los datos de “entrenamiento”.

Sería un error, en este sentido, pensar que si un método funciona bien, hace su papel, (por ejemplo: clasifica bien) al aplicarlo a los datos con los que ha sido generado el modelo obtenido es “bueno”.

Por ello es habitual reservar una parte de la información disponible para la evaluación del modelo y utilizar otra parte para generarlo. Si el volumen de datos lo permite se suele separar un conjunto de datos de **entrenamiento** (**training set**) y reservar el resto para la **prueba** (**test set**).

2.3. Validación e interpretación. Procedimientos de validación

Pero, antes de seguir, ¿qué es funcionar bien?.
Si estamos ante una tarea predictiva funcionar bien es hacer buenas predicciones. Esto nos lleva a considerar:

- Si se trata de una clasificación, el número o porcentaje de aciertos, la ratio entre aciertos y fallos, las asignaciones correctas a cada posible clase y otros indicadores similares nos permitirán evaluar la calidad de la clasificación. Muy posiblemente los errores de clasificación supongan un coste diferente dependiente de la clase de asignación errónea y, entonces será adecuado elaborar indicadores de calidad que consideren esta cuestión
- Si se trata de una regresión (predicción numérica) medidas ya conocidas como el error cuadrático medio u otras que también consideren funciones de pérdidas asociadas pueden ser de utilidad.

Si la tarea es descriptiva evaluar la calidad de su resultado puede ser , a veces , más complejo, especialmente en el caso de la obtención de reglas de asociación y en otras tareas. En el caso del clustering la relación entre la homogeneidad interna y la heterogeneidad externa de los grupos obtenidos puede ser una buena herramienta y , en este sentido la ratio entre las varianzas internas y externa, las medidas de información u otras similares serán habituales.

Pero en la medida en que el objetivo es obtener buenos resultados generalizables más allá de los datos utilizados para la obtención del modelo (conjunto de datos de entrenamiento) es importante evaluar los indicadores de bondad de la predicción, clasificación , agrupamiento etc. para otros datos no utilizados en el entrenamiento del modelo (test set)

En algunas ocasiones el número de datos disponibles es muy grande y nos podemos “permitir el lujo” de utilizar sólo una parte de ellos (un 60%, un 70% de ellos) para entrenar y obtener el modelo y reservar el resto para probar su ejecución y evaluar su funcionamiento. (2/3 y 1/3 para training y test es una proporción bastante frecuente)

Sin embargo, en otras muchas ocasiones proceder de esta manera no es posible (porque no se disponen de muchos datos) o no es razonable porque un buen ajuste del modelo necesita un cantidad mayor de datos, o la variabilidad de la base de datos es tal que aconseja considerarlos todos ya que una parte de los mismos podría sesgar esa variabilidad y dejar de ser una extracción representativa. En estos casos se suele

optar por la validación cruzada, la validación bootstrap o, en menor medida por la validación dejando uno fuera.
La validación cruzada (de k capas o particiones), k-folds cross validation, en inglés, consiste en dividir el conjunto de datos en k capas o particiones obtenidas aleatoriamente y utilizar k-1 de estas capas para entrenar el modelo y la restante para testarlo, repitiendo el proceso k veces de forma que al final se han usado todos los datos para entrenar el modelo y para testar pero nunca se han usado los datos utilizados para generar en el modelo en el proceso de evaluación. Finalmente los indicadores de calidad se obtienen promediando los indicadores de cada una de las k pruebas (tests)

La validación por **bootstraps** consiste en seleccionar por muestreo **con reemplazamiento** una cierta cantidad de datos , digamos N, Este conjunto de datos que podría contener repeticiones , se utilizará como conjunto de entrenamiento y el resto que vienen a ser 0.368 veces N, los usaremos como conjunto de prueba.
La medidas de error (o de calidad) serán una suma ponderada recíprocamente de las medidas obtenidas en el conjunto de entrenamiento y en el conjunto test:

Por ejemplo:
$\text{ERROR}_{(\text{GLOBAL ESTIMADO})} = 0{,}368\ \text{ERROR}_{(\text{ENTRENAMIENTO})} + 0{,}632\text{ERROR}_{(\text{TEST})}$

Respecto al tamaño aproximado del conjunto de prueba (test) es sencillo comprobar que es aproximadamente 0,368 veces N

La probabilidad de que un dato individal concrerto no resulte extraído en N extracciones es: 1-1/N . De forma que la probabilidad de que no resulte extraído en ninguna de la N será:

$\left(1-\frac{1}{N}\right)^N$ que si el número de extracciones es lo

suficientemente grande tenderá a $\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N = \frac{1}{e} = 0.368$

Por lo que hace a la **interpretación** , el otro constituyente de la fase del proceso KDD (Interpretación y evaluación) es evidente que un elemento de calidad exigible que todo modelo obtenido por minería de datos es que puede ser fácilmente interpretable por los expertos y/o por los usuarios finales . Unos métodos u otros tendrán de origen una más fácil o más compleja interpretabilidad que habrá que ponderar a la hora de optar por unos u otros. Un árbol de decisión o un clasificador basado en reglas es mucho más fácilmente interpretable que una red neuronal, por ejemplo.

También la interpretabilidad depende, en parte, de la tarea de que se trate. El resultado final de un patrón de agrupación es una estructura de grupos que es auto-explicada como tal, si bien , a menudo, su justificación o comprensión requerirá de cierto análisis posterior

2.4. Expresividad: Ajuste, sobreajuste y subajuste

El análisis detallado de la expresividad de una determinada técnica o el estudio de la expresividad de los distintos tipos de modelos es un asunto muy complejo que desborda la pretensiones del curso. Sin embargo, cierta consideración sí debe dársele

ya que es una cuestión que afecta o puede afectar, en buena medida, a la calidad de los resultados que puedan obtenerse con cada modelo .
La expresividad de un modelo es la capacidad que tiene de dar cuenta de los datos que se usan, se han usado o se van a usar para construirlo.

A pesar de lo que pudiera pensarse en una primera aproximación ingenua no es demasiado conveniente que el modelo “exprese” demasiado bien los datos que se han usado para construirlo. En efecto, si existe un excesivo ajuste del modelo a los datos (overfitting= sobreajuste), el modelo no podrá dar cuenta de otros datos futuros , perderá generalidad a la hora de aplicarse a otros conjuntos de datos. Pensemos en un árbol de decisión que da como resultado tantas ramas como datos se han presentado. Predecirá a la perfección el conjunto de datos suministrado, pero no tendrá ninguna utilidad , más allá de eso.

El subajuste (underfitting) es el problema contrario, el modelo obtenido es demasiado genérico, no es capaz ni siquiera de dar cuenta de la información muestral suministrada, y no tendrá ninguna utilidad. Un posible ejemplo podemos encontrar en cualquier modelo de predicción de pésimo ajuste una regresión lineal sobre un conjunto de datos de clara variación exponencial, por ejemplo.

Obviamente un buen modelo de minería de datos debe ajustarse bien pero sin caer en el sobreajuste y esa es la medida adecuada de expresividad.

Referencias:

Hernández, J., Ramirez, M, & Ferri, C. (2010). *Introducción a la minería de datos.* Pearson Prentice Hall.

Mitchell, T. (1977). *Machine Learning.* McGraw Hill.

Solomonoff, R. (1966). A fromal theory of inductive inference, Part I. *Information And Control,PartI, Vol7,No1*, 1-22.

Wolpert, D., & Macready, W. (1997). No free lunch theorems for search. *IEE Transactions on Evolutionary Computation,*, (págs. 67-82). Santa Fe.