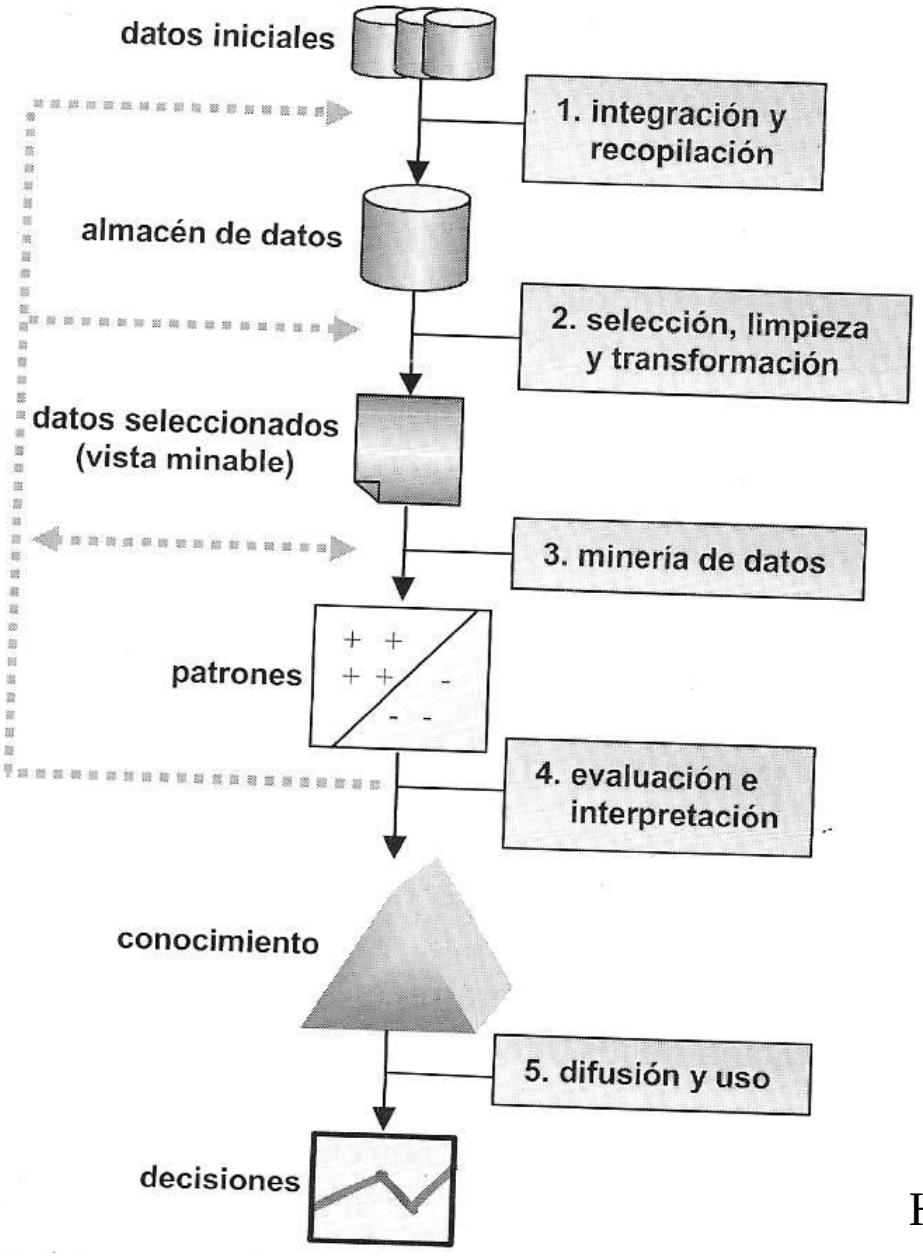
Minería de datos

Unidad 2. El proceso KDD

Evaluación, difusión y uso

M en I Sara Vera Noguez

El proceso KDD



Hernández, 2008 p.20

La parte iterativa

- Una vez obtenido el modelo se debe evaluar
- Si satisface las necesidades, no es necesario hacer iteraciones; en otro caso si debe hacer otra iteración.
- Las iteraciones pueden ser para modificar el algoritmo, la técnica, afinar los datos o incluso los requerimientos.

Fase de evaluación e interpretación

Se busca que los patrones descubiertos tengan las siguientes cualidades:

- Ser precisos
- Ser comprensibles
- Ser interesantes

Técnicas de evaluación

- Validación con datos conocidos, pero distintos a los usados en el entrenamiento.
- **Validación simple**: dividir el conjunto de datos en dos: un conjunto de entrenamiento y uno pequeño de prueba, los de prueba no se deben usar en el entrenamiento.

Técnicas de evaluación

- **Validación cruzada**: cuando se tiene pocos datos, el conjunto se divide en dos: A y B; la división es aleatoria y los conjuntos son del mismo tamaño, primero se entrena el modelo con los datos de A y se validan con B, se calcula el error; se entrenan los datos con B, se validan con A y se calcula el error, finalmente se entrenan con ambos, y se usa el modelo con el menor error

Técnicas de evaluación

- **Validación cruzada con n pliegues**: Es una variación de la cruzada, y consiste en divir el conjunto en n grupos; se usa uno para el entrenamiento y n-1 para pruebas, se calcula el error y se repite el proceso n veces cambiando el conjunto de entrenamientos.

¿y como medimos si funciona?

- Dependiendo del problema.
- Para cada problema tenemos una medida de evaluación.

Medidas de evaluación

- Dependiendo del problema, **para clasificación** se mide la **precisión predictiva**, $p=c/n$
- c es el número de instancias del conjunto de prueba clasificadas correctas
- n es el número de instancias totales
- Se busca maximizar el valor de p tanto para el conjunto de entrenamiento como **para el de prueba**

Medidas de evaluación

- **Para reglas de asociación**, se evalúa de forma separada cada regla, considerando:
- **Cobertura** o soprte = #de instancias a las que la regla aplica y predice correctamente
- **Confianza**: proporción de instancias que la regla predice correctamente,= cobertura/# de reglas a las que se puede aplicar

Medidas de evaluación

- **Para regresión**, se evalúa el **error cuadrático medio** calculado como el prodio de los cuadrados de los valores esperados- valores calculados

Medidas de evaluación

- **Para agrupamiento**, la medida aplicada dependerá del método utilizado para su implementación, que de forma general miden la cohesión de los grupos, como por ejemplo la distancia media al centro, la **distancia** media entre grupos; o bien la **densidad**

Relación tarea -medida

- En equipos de dos integrante complemente la tabla de tareas de minería especificando la medida a evaluar en cada tarea.

Interpretación y contextualización

- Las medidas de evaluación debe ser llevadas al contexto específico, ya que este influye en la evaluación del modelo e interpretación de los resultados
- p.e. para un problema de clasificación con distribuciones de clases no balanceas un valor alto de precisiones no es sinónimo de eficiencia, (si una clase acumula el 90% de la población, y la precisión es .9, pude no saber clasificar a los elementos de clase minoritaria)

Interpretación y contextualización

- Para contemplar otros aspectos, se pueden usar herramientas como la **matriz de costos** de clasificación que es un tipos especial de **matriz de confusión**.
- Cuando no es viable estimar los costos de los errores se pude aplicar el **análisis ROC** (Receiver Operating Characteristic)

Fase de difusión y uso

- El uso de los modelos (generados y validados) se enfoca en dos escenarios:
- Para recomendar acciones con base en los resultados del modelo (ya se usado por una persona o un sistema)
- Para aplicar el modelo en distintos conjuntos de datos.

Fase de difusión y uso

- Es necesaria la difusión de los resultados obtenidos entre los posibles usuarios.
- Y dar seguimiento para la evolución del modelo, evaluándolo periódicamente ya que los patrones pueden evolucionar.
- Se pude requerir un ajuste o reconstrucción del modelo

Bibliografía básica

- Hernández Orallo, J., M. J. Ramírez Quintana, et al. (2004). Introducción a la Mineria de Datos. España, Pearson Educación SA.
- Han, D. J. (2007). Principles of Data Mining, MIT Press.
- Maimon, O. Z. and L. Rokach (2005). Data mining and knowledge discovery handbook. USA, Springer.
- Pérez López, C. and D. Santín Gonzalez (2006). Data Mining-Soluciones Con Enterprise Miner. México, Alfaomega, Ra-Ma.

Bibliografía complementaria

- Everitt, B.S. (1994). A Handbook of Statistical Analyses using S-Plus. Chapman and Hall.
- Inmon, W.H. (1996). Building the Datawarehouse. J.Wiley & Sons.
- Han, J. and M. Kamber (2006). Data mining: concepts and techniques, Morgan Kaufmann.
- Kimball, R (1996). The Data Warehouse Toolkit. John Wiley & Sons.
- Hastie, T., R. Tibshirani, et al. (2005). The elements of statistical learning: data mining, inference and