

## Hacia la consideración de aspectos de calidad de datos en procesos de minería: el caso de las técnicas de clasificación

Roberto Espinosa  
Dept. de Informática  
Universidad de  
Matanzas, Cuba  
respinosa@umcc.cu

Jose Zubcoff  
Dept. de Ciencias del  
Mar y Biología Aplicada  
Universidad de Alicante,  
España  
jose.zubcoff@ua.es

Marta Zorrilla  
Dept. de Matemáticas,  
Estadística y Computación  
Universidad de Cantabria,  
España  
marta.zorrilla@unican.es

Jose-Norberto Mazón  
Dept. Lenguajes y  
Sistemas Informáticos  
Universidad de Alicante,  
España  
jnmazon@dlsi.ua.es

### Resumen

El éxito en la búsqueda de conocimiento a partir de grandes cantidades de datos radica en la calidad de los mismos. Hasta ahora los aspectos de calidad de los datos se han enfocado principalmente a la limpieza de los datos: detección de duplicados, valores atípicos, perdidos, incompletos o conflictos en instancias, entre otros. En este trabajo se presenta un caso de estudio que nos ha permitido determinar ciertos aspectos de calidad que pueden mejorar la expectativa de éxito en el análisis evitando resultados erróneos, incorrectos o poco fiables. Este es un primer paso hacia la consideración de manera sistemática y estructurada de criterios de calidad específicos para minería de datos que ayude al minero de datos en sus objetivos.

### 1. Introducción

La minería de datos es un proceso clave en las aplicaciones de inteligencia de negocio ya que permite extraer conocimiento útil a partir de un conjunto de datos. Un proceso típico de minería de datos [16] empieza con un conjunto de datos, del cual el analista selecciona un subconjunto que va a formar parte del análisis. Además, debe seleccionar la técnica y más concretamente el algoritmo a aplicar a ese subconjunto de datos. Éste se aplica con unos parámetros en función del objetivo buscado o bien con los parámetros por defecto. Finalmente, se obtienen los patrones de comportamiento común en los datos y el análisis de estos patrones es lo que permite descubrir conocimiento útil en los datos [17].

El éxito de este proceso de minería de datos depende en gran medida de la calidad de los datos implicados [18]. Errores de introducción, valores inadecuados, atípicos, perdidos, u otros errores en los datos pueden hacer que el patrón resultante no sea correcto. Sin embargo, la limpieza de los datos no es el único factor de la calidad de datos que puede afectar negativamente al proceso de minería, existiendo otros factores menos obvios a primera vista pero igual de relevantes para la minería de datos. Para considerar cada uno de estos factores, se debe tener en cuenta que una definición típica de calidad de datos es “adecuación al uso” [1], lo que en el contexto de la minería de datos significara que los datos que participen en el proceso tengan en cuenta el contexto, esto es, el conocimiento del problema que se está tratando para que el usuario pueda obtener conocimiento útil al aplicar técnicas de minería de datos. Así, se podría evitar que la aplicación de dichas técnicas resulte en conocimiento superfluo, contradictorio o incluso erróneo.

Existen un buen número de dimensiones de calidad de datos (ver [2], [3] y el estándar ISO [4]) que se podrían tener en cuenta en el proceso de minería de datos como por ejemplo en la selección adecuada de los atributos que formarán el modelo de minería de datos, o la correcta selección de parámetros y/o algoritmos a utilizar en el proceso de minería de datos. Resulta conveniente que tales aspectos de calidad sean tenidos en cuenta en etapas previas al proceso de minería de datos, y en la medida de lo posible, en la fase de requisitos de usuario. Por ejemplo, cuando se trabaja con grandes cantidades de datos (intensional y extensionalmente), un minero de datos no experto y no consciente del contexto, puede construir un modelo seleccionando un conjunto de atributos y una técnica determinada, pero esto no asegura que

el patrón que obtenga sea correcto aunque los datos estén limpios y libres de errores.

Por tanto, se debe determinar aquellos aspectos de calidad de datos que puedan afectar ampliamente al resultado de las técnicas de minería de datos, pudiéndose detectar y corregir en etapas tempranas de diseño. Concretamente, en este artículo se presenta un caso de estudio que ha permitido determinar algunos aspectos a tener en cuenta cuando se aplican técnicas de clasificación: desde la detección de valores atípicos, datos desbalanceados o altamente desbalanceados, datos que se encuentran en una o varias jerarquías de conceptos entre otros. Esto nos permitirá detectar si los datos de origen son adecuados para llevar a cabo las técnicas de minería de datos deseadas o si, por el contrario, existen restricciones marcadas por el contexto o los propios datos que deben ser tenidas en cuenta con el fin de extraer el conocimiento útil.

En la Sección 2 se comentan los trabajos relacionados, en la Sección 3 se desarrolla el caso de estudio, en la Sección 4 se detallan los criterios de calidad para técnicas de clasificación y finalmente en la Sección 5 se presentan las principales conclusiones del trabajo.

## 2. Trabajos relacionados

Existen numerosos trabajos en la literatura que abordan el problema de calidad de los datos desde el punto de vista de la limpieza de los mismos: i) para la detección de duplicados [9], eliminación de los mismos [10][13], reconocimiento de instancias bajo distintas etiquetas [12], comparaciones de cadenas de caracteres, etc., ii) resolución de conflictos en instancias [23] usando técnicas específicas de limpieza [22], iii) valores atípicos [21], perdidos o incompletos [20], entre otros. La estandarización de los datos también ha sido considerada por varias técnicas de limpieza para resolver problemas como la estructura heterogénea de los datos (*i.e.* representación estándar de fechas) [19]. Dentro del ámbito de la calidad de datos para procesos de minería de datos, y hasta donde hemos podido comprobar en las fuentes bibliográficas, no se establecen mecanismos para definir desde etapas tempranas del proceso KDD ciertos criterios de calidad que resultan sumamente importantes para obtener resultados adecuados.

## 3. Caso de estudio

A continuación se presenta el caso de estudio que nos permite mostrar algunos aspectos a tener en cuenta en la selección de los datos para poder usar de manera satisfactoria técnicas de clasificación.

Los datos de nuestro caso de estudio contienen información estadística del comportamiento de jugadores en partidos pertenecientes a varios torneos de baloncesto (masculino y femenino), de donde se recogen, por cada partido y jugador un conjunto de indicadores (Tabla 1). Además, se tienen en cuenta datos de los propios jugadores como sexo, altura, peso, etc.

Tabla 1. Conjunto de indicadores estadísticos.

Indicadores	Nombre	Abreviatura
Defensivos	Bolas ganadas.	BG
	Rebotes defensivos.	RD
	Fallar en el enfrentamiento a un adversario que penetra hacia el cesto.	FE
	No recuperar el rebote tras el lanzamiento del equipo rival.	NR
	Asistencias defensivas.	AD
Ofensivos	Asistencias.	A
	Pérdidas de balón.	PB
	Tiros libres errados.	TLE
	Tiros errados de 2.	TE2
	Tiros errados de 3.	TE3
	Rebotes ofensivos.	RO
	Puntos por Tiros Libres Anotados.	PA1
	Puntos anotados de 2.	PA2
Puntos anotados de 3.	PA3	

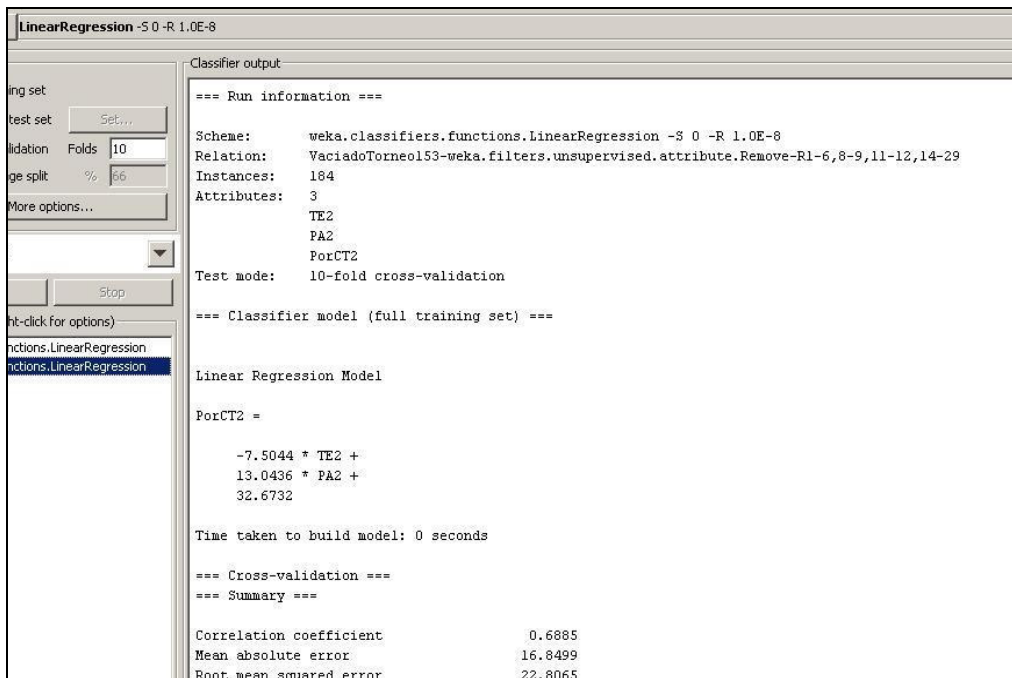


Figura 1. Resultados obtenidos al aplicar Regresión Lineal entre TE2, PA2 y PorT2.

Uso común	Permanencia en la cancha	PC
-----------	--------------------------	----

Tabla 2. Posiciones de los jugadores

Posición	Abreviatura	Responsabilidad principal
Base	B	Organizar el juego y asistir a sus compañeros.
Alero	A	Anotar canastas. Llevar peso ofensivo del equipo.
Pívot	P	Rebotear. Garantizar el juego bajo canasta.

Para cada jugador se conoce su posición (Tabla 2), y una evaluación global nominal (Poco Integral, Integral y Muy Integral) que se establece

con una fórmula que evalúa su actuación en cada partido de acuerdo a su posición en el juego y los indicadores ofensivos y defensivos.

En este caso de estudio se utiliza este conjunto de datos para generar varios modelos de clasificación y demostrar cómo una selección no adecuada de atributos lleva a obtener un patrón inútil, poco preciso y carente de valor.

Nuestro objetivo, por tanto, es señalar los principales aspectos de calidad de datos, en su sentido más amplio, que afectan a los resultados cuando no se tiene en cuenta el conocimiento del dominio del problema (el experto en baloncesto en este caso). De esta manera podremos determinar qué criterios se han de contemplar a la hora de seleccionar los atributos para cada determinada técnica con objeto de poder notificar al minero de datos no experto en el dominio que el proceso KDD definido puede presentar conflictos de calidad. A continuación se exponen varias situaciones donde se obtienen resultados no útiles, no fiables o incluso inconsistentes, como consecuencia de aplicar algoritmos de clasificación sobre un conjunto de atributos no adecuado:



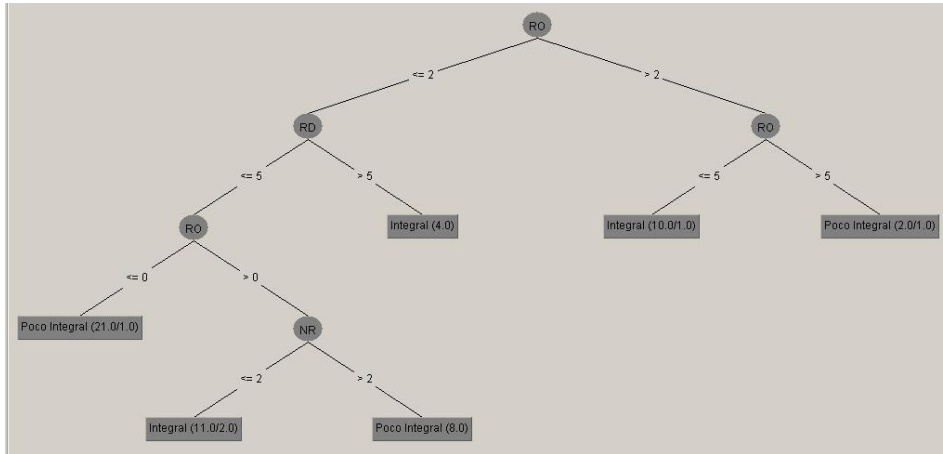


Figura 4. Resultados filtrados para los jugadores cuya posición es Pivot.

presenta el ejemplo en el que se construye un clasificador para determinar la evaluación global de los jugadores. Para ello se utiliza los atributos de la tabla 1. Como se puede observar en la Fig.3, con RD (Rebotes Defensivos) > 5, ya el jugador es clasificado de Integral, el 90% de las veces. Pero resulta que esta clasificación no ha tenido en cuenta la posición del jugador, dato que resulta determinante para la evaluación global del mismo de acuerdo al conocimiento del experto.

Si además a esto se le añade el hecho de que el experto conoce que existen ciertos indicadores que son más afines a cada posición, se podría sugerir al minero de datos la generación de tres conjuntos de datos, uno para cada posición con sus indicadores correspondientes, lo que redundará en un conseguir un clasificador más preciso.

Al aplicar nuevamente el algoritmo de clasificación por separado para cada posición en la cancha (Base, Alero y Pivot), los resultados que se obtuvieron evidenciaron que se les otorga mayor

peso a los indicadores propios según la función dentro del juego de cada jugador. Para los jugadores cuya posición en el campo es Pivot se identifica como el indicador de más peso los Rebotes Ofensivos (RO) (Fig. 4), para los Aleros los Puntos Anotados de 2 (PA2) (Fig. 5), y para los Base las Asistencias (A) (Fig. 6).

Los resultados que se presentan en este caso de estudio evidencian que el conocimiento del dominio del problema así como el significado y modo de cálculo, si existe, de los atributos es imprescindible para generar modelos de minería útiles y exitosos.

Esto nos lleva a pensar que si esta información puede quedar recogida en el conjunto de algún modo, se podría construir una herramienta que ayudara al modelado de procesos KDD y se evitara así la extracción de conocimiento erróneo o inútil que lleve a la toma de decisiones equivocada.

Los aspectos que afectan a la calidad de los datos en técnicas de clasificación se comentarán en detalle en la siguiente sección.

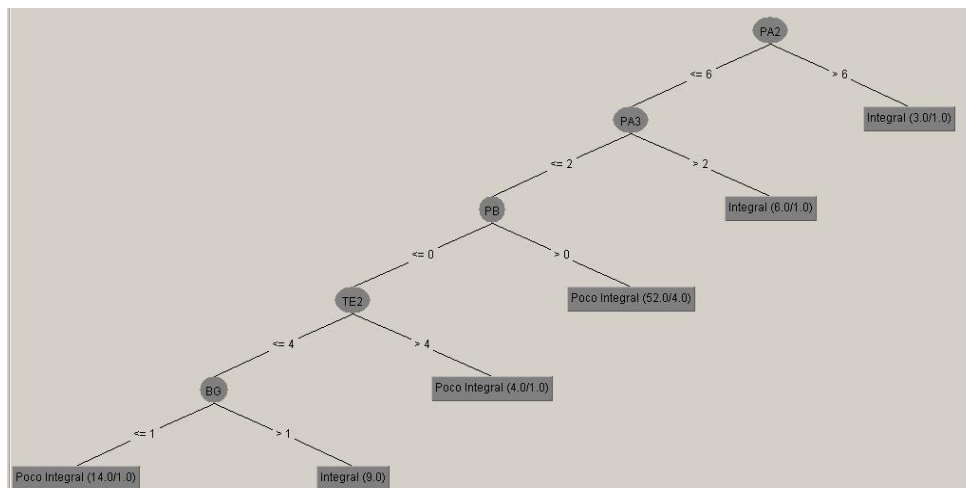


Figura 5. Resultados obtenidos para los jugadores cuya posición es Alero.

Como en todo proceso de clasificación, la calidad de los datos así como otras restricciones de dominio, han determinado la posibilidad de encontrar patrones novedosos y que aporten información útil. Mediante el caso de estudio se ha detectado la necesidad de ayudar al analista en el proceso de comprobar correlaciones en los datos seleccionados, comprobar si estaban desbalanceados y verificar si algunos clasificadores presentaban diferentes patrones en función de algún atributo y, por tanto, tampoco aporta información novedosa (por ejemplo si el jugador es base, alero o pivót tendrá diferentes patrones en rebotes, tiros acertados, etc.). Todo esto, además de limpiar los datos de errores, valores nulos, etc.

#### 4. Aspectos de calidad

A partir de este caso de estudio hemos detectado tres aspectos relacionados con la calidad de datos en su “adecuación al uso” de técnicas de clasificación que pueden y deben ser abordados en etapas tempranas del proceso de extracción del conocimiento. En concreto los datos seleccionados para la aplicación de una técnica de clasificación deberían i) evitar datos correlacionados, ii) evitar datos altamente desbalanceados y finalmente, iii) seleccionar datos según el contexto del problema..

Los datos correlacionados afectan al patrón resultante de manera tal que aumentan la complejidad del mismo sin aportar información novedosa.

Por otro lado, los datos altamente desbalanceados dan como resultados patrones que

no son fiables porque representan un sobreajuste a los datos correspondientes al criterio que aporta mayor cantidad de datos.

En cuanto a los datos que no están seleccionados en base a un criterio del dominio, pueden dar como resultado patrones más complejos y con poca o nula información novedosa (según hemos mostrado en nuestro caso de estudio).

Si se tienen en cuenta estos criterios de calidad en etapas tempranas del desarrollo de un proyecto de minería de datos, se podría determinar si es adecuado o no aplicar ciertas técnicas de minería de datos para obtener conocimiento útil, de tal manera que se evitara la aplicación de técnicas que resulten en conocimiento superfluo, contradictorio o incluso erróneo. Es decir, además del proceso habitual de limpieza de datos se debería tener en cuenta de manera explícita en una etapa de análisis de requisitos la calidad de los datos de los que se dispone según estos tres aspectos de calidad adicionales encontrados para la aplicación de técnicas de clasificación:

- Evitar que el usuario seleccione algunos atributos que estén directamente relacionados con la clase. Por ejemplo, evitando seleccionar atributos a distinto nivel en una jerarquía, detectando dependencias funcionales del esquema donde se encuentran los datos o aplicando técnicas de detección de correlaciones como los algoritmos de regresión sobre los atributos del modelo.
- Alertar sobre deficiencias en la calidad de los datos en dichas etapas tempranas de diseño. Esto es, si los datos están desbalanceados, analizando el subconjunto de datos

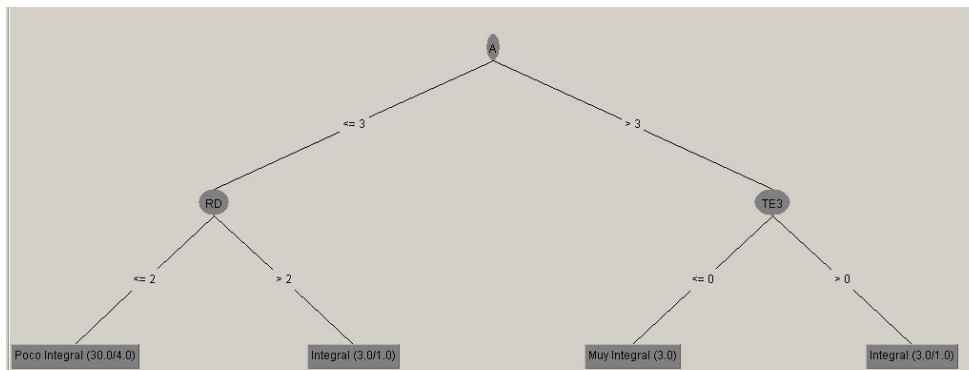


Figura 6. Resultados filtrados para los jugadores cuya posición es Base.

seleccionado en el modelo conceptual de minería de datos.

- Respecto a la selección de atributos y filtrado de instancias, partir de criterios del dominio puede hacerse de una manera semiautomática en las etapas de modelado conceptual de la minería de datos, usando criterios como pre-selección de etiquetas automáticamente a partir de los datos desbalanceados o bien con la ayuda del experto en el dominio seleccionando las etiquetas a partir de atributos que ya se conozcan sus patrones de comportamiento (en nuestro caso de estudio se sabía a priori que los patrones resultantes iban a ser distintos en función de la posición del jugador: base, alero o pivot).

Considerar la calidad de los datos disponibles permitirá conocer en la etapa de análisis de requisitos qué algoritmos se pueden aplicar o cuáles darán mejores resultados. Por lo tanto, se tendrán mejores oportunidades de alcanzar el éxito en el proceso de minería de datos con técnicas de clasificación.

## 5. Conclusiones

En este artículo se ha presentado un caso de estudio que nos ha permitido determinar ciertos aspectos de calidad que deben poseer los datos para llevar a cabo el proceso de minería de datos con técnicas de clasificación. Cumpliendo con estos criterios, se podrá obtener conocimiento útil al aplicar las técnicas de minería de datos requeridas por el usuario y evitar que la aplicación de dichas técnicas resulte en conocimiento superfluo, contradictorio o incluso erróneo.

Nuestro trabajo futuro consiste en la aplicación de estos aspectos de manera formal, sistemática y estructurada por medio de un proceso de diseño de minería de datos dirigido por los requisitos de usuario. A grandes rasgos, a partir de los requisitos de minería de datos se especificará un modelo multidimensional a nivel conceptual [6][7] con el fin de estructurar los datos de manera eficaz para facilitar la tarea de modelado de procesos de minería de datos [5][8]. Este modelo de minería de datos determinado por los requisitos de usuario debe ser reconciliado con las fuentes de datos disponibles según los criterios de calidad de datos correspondientes a la técnica de minería de datos elegida con el fin de que la

extracción de conocimiento satisfaga al usuario a la vez que sea coherente con los datos disponibles.

## Referencias

- [1] Ge M. and Helfert M. (2007), Develop a research agenda: a review of information quality research, 12th International Conference on Information Quality, Boston, Massachusetts, USA.
- [2] Brodie, M. L. (1980), Data quality in information systems. *Information and Management*, 3(6), pp. 245-258.
- [3] Berti-Equille, L. (2007). *Quality Awareness for Managing and Mining Data*. PhD. Dissertation. University of Rennes, France.
- [4] ISO-25012, ISO/IEC 25012: Software Engineering-Software product Quality Requirements and Evaluation (SQuaRE)- Data Quality Model. 2008.
- [5] Zubcoff, J., Pardillo, J. and Trujillo, J. (2008), Integrating the development of data mining and data warehouses via model-driven engineering, *Proceedings of Apoyo a la Decisión en la Ingeniería del Software*, ADIS. Oviedo, Spain.
- [6] Mazon, J. and Trujillo, J. (2007). A model driven modernization approach for automatically deriving multidimensional models in data warehouses. In *ER*, pp. 56–71, 2007.
- [7] Mazon, J. and Trujillo, J. (2008). An MDA approach for the development of data warehouses. *Decis. Support Syst.*, 45(1):41–58.
- [8] Zubcoff, J. and Trujillo, J. (2006) *Conceptual Modeling for Classification Mining in Data Warehouses*, *Proceedings of International Conference on Data Warehousing and Knowledge Discovery, DaWaK*, pp. 566–575.
- [9] Elmagarmid, A., Ipeirotis, P. and Verykios, V. (2007). Duplicate record detection: A survey. *TKDE*, 19(1):1–16.
- [10] Ananthakrishna, R., Chaudhuri, S. and Ganti, V. (2002). "Eliminating Fuzzy Duplicates in Data Warehouses," *Proc. 28th Int'l Conf. Very Large Databases (VLDB '02)*.
- [11] Little RJ, Rubin DB (1987) *Statistical analysis with missing data*. Wiley, New York

- [12] Lim L, Srivastava J, Prabhakar S, Richardson J (1993) Entity identification in database integration. In: Proceedings of the 9th international conference on data engineering (ICDE), Vienna, Austria, pp 294–301
- [13] Missier, P., Batini C. (2003). A multi-dimensional model for information quality in CIS. In: Proceedings of the 8th international conference on information quality (IQ), MIT, Cambridge, MA, USA.
- [14] Sarawagi, S. eds. (2000) Special issue on data cleaning. IEEE Data Engineering Bulletin, 23(4).
- [15] Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London
- [16] Frawley, W. Piatetsky-Shapiro, G., Matheus, C. (1992) Knowledge discovery in databases: an overview, AI Magazine 13, pp.213– 228.
- [17] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.: (1996), The KDD process for extracting useful knowledge from volumes of data, Commun. ACM 39(11), 27–34.
- [18] González-Aranda, P., Menasalvas, E., Millán, S., Ruiz, C. and Segovia, J.: (2008). Towards a Methodology for Data Mining Project Development: The Importance of Abstraction, Data Mining: Foundations and Practice, pp. 165–178.
- [19] Sarawagi, S. (2000). Special issue on data cleaning. IEEE Data Engineering Bulletin, 23(4).
- [20] Theodoratos, D., Bouzeghoub, M. (2001). Data currency quality satisfaction in the design of a data warehouse. Special Issue on design and management of data warehouses. Int J Coop Inf Syst 10(3):299–326.
- [21] Knorr, E., Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th international conference on very large data bases (VLDB), New York City, USA, pp 392–403.
- [22] Rahm, E., Do, H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 23(4):3–13.
- [23] Pearson, R.K. (2002). Data mining in face of contaminated and incomplete records. In: Proceedings of SIAM international conference on data mining.