Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones

Programas de Maestría y Doctorado en Ingeniería Telemática
Seminario de Investigación

Evaluación, limpieza y construcción de los datos: un enfoque desde la inteligencia artificial

David Camilo Corrales Muñoz
Estudiante de Doctorado
05 de diciembre de 2014

1. Introducción

El propósito de la relatoría es presentar los avances de la propuesta de doctorado alrededor del tema de calidad de los datos en tareas de aprendizaje supervisado. Para esto se abordaron los siguientes puntos: una breve introducción que resalta los principales conceptos para entender la propuesta de investigación, la motivación para el desarrollo del proyecto, acompañado de una visión general de los trabajos relacionados, la pregunta de investigación y los objetivos preliminares. A continuación se explicará de manera breve cada uno de los puntos abordados en la relatoría.

2. Introducción

Inicialmente fueron presentados ante la audiencia los siguientes conceptos:

2.1. Aprendizaje supervisado: definido como el proceso en el cual un algoritmo aprende a partir del conjunto de ejemplos (datos de entrenamiento) con la intensión de predecir o clasificar un nuevo dato de entrada [1]**.**

2.2. Definiciones de framework para calidad de datos: se llevan a colación las siguientes definiciones de framework para la calidad de datos:

- Herramienta para evaluar la calidad de los datos dentro de una organización [2].
- Define un modelo de su entorno de datos, identificando los atributos de calidad de datos adecuados, los cuales son analizados en un contexto actual o futuro, con la finalidad de guiar a la mejora de la calidad de datos [3].
- Un framework debe evaluar, y proporcionar un esquema para analizar y resolver problemas de calidad de datos para una gestión proactiva [4].

- Buscan evaluar áreas donde los procesos de baja calidad reducen la rentabilidad de una organización [5].

3. Escenario de Motivación:

En los últimos años, son innumerables las tareas de clasificación, detección y predicción, que han sido automatizadas a través de la aplicación de algoritmos de aprendizaje automático (AA). Estos algoritmos requieren de un conjunto de datos durante la fase de aprendizaje (proceso de entrenamiento) con el propósito de llevar a cabo la clasificación, detección o predicción, sin embargo los algoritmos de AA carecen de mecanismos que garanticen la calidad de la información. La merma en la calidad de los datos se debe, principalmente, a errores cometidos durante el proceso de captura de los datos lo que lleva a la obtención de resultados erróneos. Además, hoy en día, los datos se obtienen de diversas fuentes y son de tipos muy distintos (web, imágenes, videos, texto, sensores, etc.).

En este orden de ideas se ha detectado que los sistemas informáticos adolecen de mecanismos que evalúen la calidad de los datos con el fin de garantizar una mejor respuesta en la clasificación y predicción de tareas basado en algoritmos de inteligencia artificial.

4. Estado del arte

En esta sección fueron explicados de manera general las aproximaciones existentes, relacionadas con el problema de investigación declarado. Estos trabajos fueron organizados según las tareas: verificar la calidad de los datos, seleccionar los datos, limpiar los datos y construir los datos de las fases comprensión y preparación de los datos de la metodología CRISP-DM [6] como se puede observar en la Figura 1:

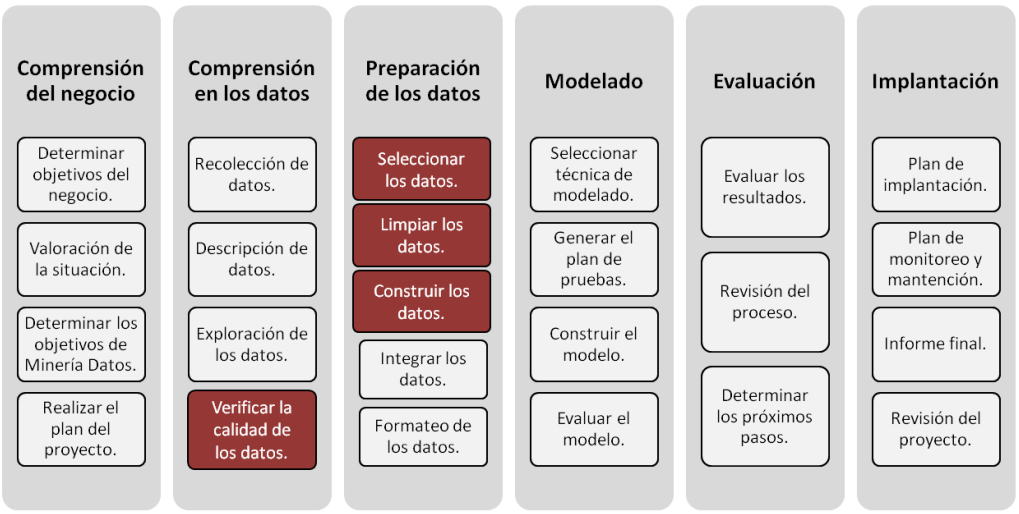

Figura 1. Fases de la metodología CRISP-DM.

A su vez, los trabajos fueron categorizados según los problemas que abordan cada tarea de la metodología CRISP-DM obteniendo 38 artículos de pérdida de datos, 34 de valores, atípicos, 18 de ruido, y 14 de inconsistencia, por otra parte para el área de reducción de la dimensionalidad aún no se ha realizado una revisión bibliográfica. En la Figura 2 se presenta un resumen.

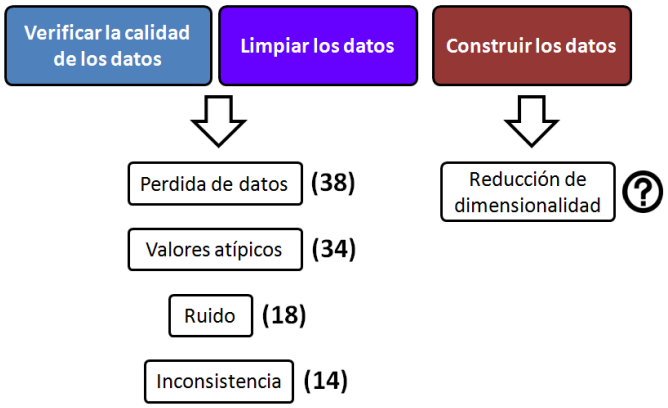

Figura 2. Clasificación de los trabajos relacionados.

De esta forma se encontraron 38 trabajos de investigación para la categoría de pérdida de datos, los cuales utilizaron los siguientes algoritmos (Figura 3):

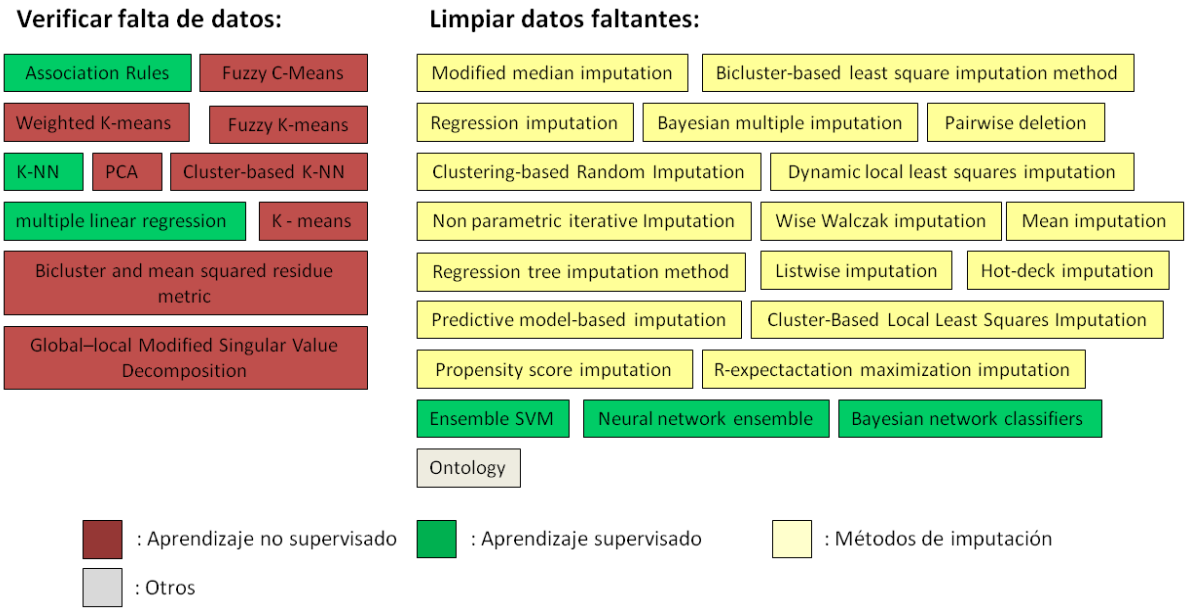

Figura 3. Algoritmos utilizados para la detección de pérdida de datos.

Los algoritmos encontrados en la Figura 3, fueron agrupados según sus características en: aprendizaje supervisado y no supervisado, métodos de imputación y otros, llegando a la conclusión que hasta el momento no existen trabajos que integren todos los problemas por los cuales se genera la perdida de datos.

Por otra parte los algoritmos utilizados detectar valores atípicos fueron los siguientes (Figura 4):

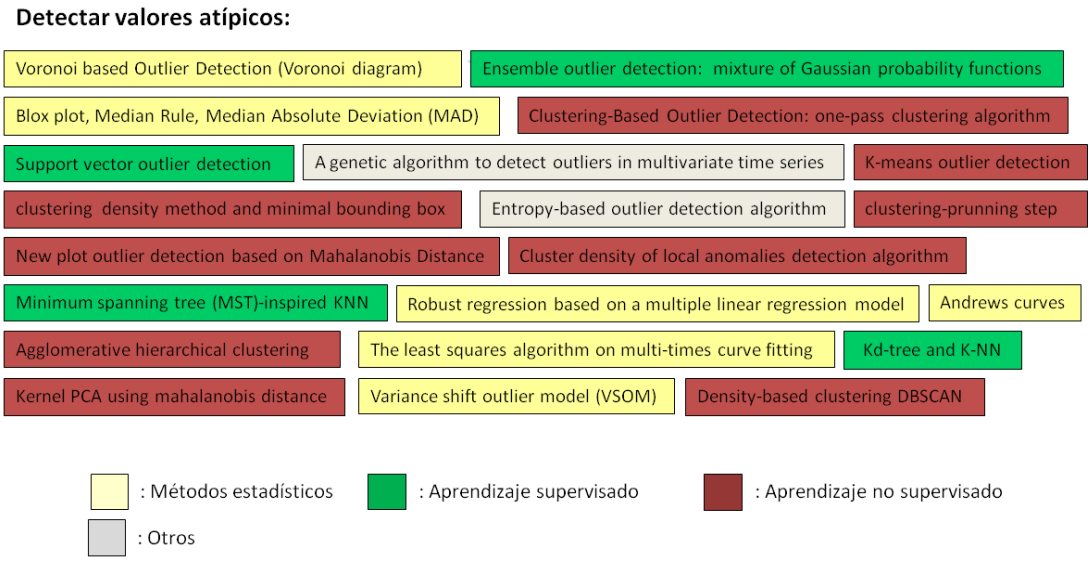

Figura 4. Algoritmos utilizados para la detección de valores atípicos.

Los algoritmos encontrados en la Figura 4, fueron agrupados según sus características en: métodos estadísticos, aprendizaje supervisado y no supervisado, y otros enfoques, llegando a la conclusión que las técnicas de aprendizaje supervisado y no supervisado se vienen trabajando desde hace un par de años hasta la actualidad, y hasta el momento no existen trabajos que integren todos los problemas de valores atípicos.

En la Figura 5, son presentados los algoritmos que intentan solucionar los problemas de ruido en un conjunto de datos.

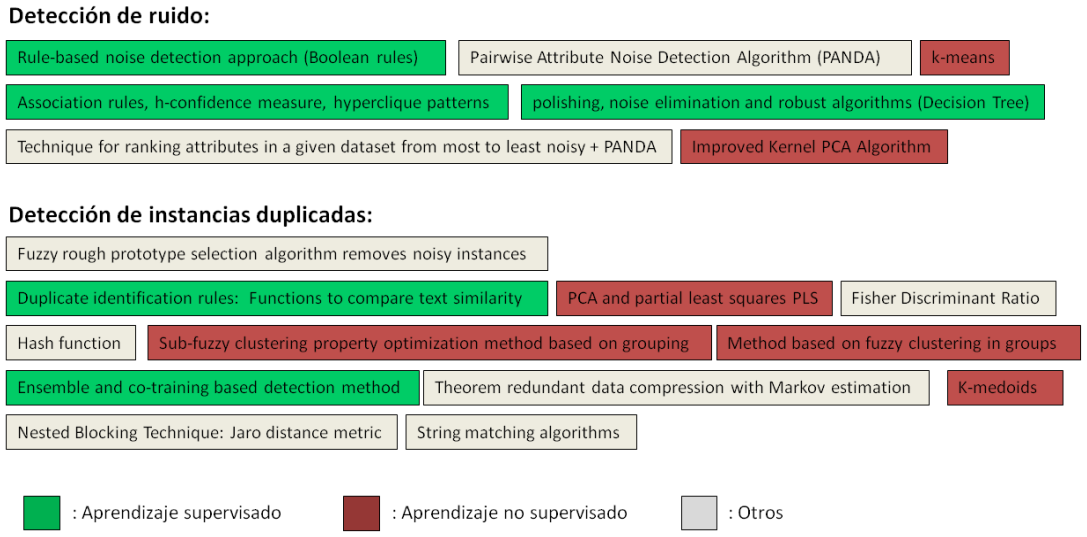

Figura 5. Algoritmos utilizados para detectar ruído.

En este sentido, los algoritmos presentados en la Figura 5, fueron clasificados según sus características en: aprendizaje supervisado y no supervisado, y otros enfoques, llegando a la conclusión que las técnicas de aprendizaje supervisado y no supervisado se vienen trabajando desde hace un par de años hasta la actualidad, y hasta el momento no existen trabajos que integren todos los problemas asociados al ruido en un conjunto de datos.

Finalmente, en la Figura 6 son presentados los algoritmos que detectan inconsistencias en un conjunto de datos:

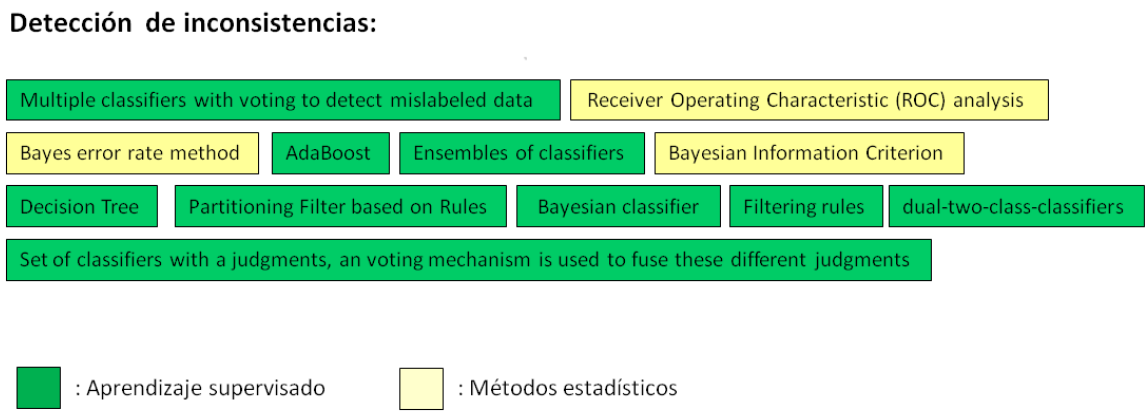

Figura 6. Algoritmos utilizados para detectar inconsistencias en los datos.

De manera similar a las categorías expuestas anteriormente, los algoritmos presentados en la Figura 6, fueron clasificados en: aprendizaje supervisado y métodos estadísticos, llegando a la conclusión que las técnicas de aprendizaje supervisado son las que más se trabajan actualmente, y hasta el momento no existen trabajos que integren todos los problemas asociados datos inconsistentes.

5. Pregunta de investigación

Con base en las consideraciones descritas anteriormente, se ha planteado la siguiente pregunta de investigación: ¿Cómo mejorar la calidad de los datos para tareas de predicción y clasificación en diferentes dominios de aplicación, a través de técnicas de inteligencia artificial?

6. Objetivos

Objetivo General

Desarrollar un framework para la evaluación de la calidad, limpieza y construcción de datos, basado en algoritmos de inteligencia artificial para tareas de clasificación y predicción en diferentes dominios de aplicación.

Objetivos Específicos

- Establecer mecanismos para la evaluación de la calidad de los datos de entrenamiento, mediante de técnicas de Inteligencia Artificial.

- Definir técnicas de limpieza en los datos de entrenamiento basado en algoritmos de Inteligencia Artificial.

- Seleccionar estrategias para reducir la dimensionalidad en los datos de entrenamiento a partir de algoritmos de Inteligencia Artificial.

- Desarrollar y evaluar experimentalmente un prototipo que valide la aproximación propuesta.

Bibliografía

[1] B. S. Araujo, *Aprendizaje automático: conceptos básicos y avanzados : aspectos prácticos utilizando el software Weka*. España: Pearson Prentice Hall, 2006.

[2] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, pp. 5-33, 1996.

[3] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, pp. 211-218, 2002.

[4] M. J. Eppler and D. Wittig, "Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years," in *IQ*, 2000, pp. 83-96.

[5] K. Kerr and T. Norris, "The Development of a Healthcare Data Quality Framework and Strategy," in *IQ*, 2004, pp. 218-233.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, *et al.*, "CRISP-DM 1.0 Step-by-step data mining guide," 2000.