Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones

Programas de Maestría y Doctorado en Ingeniería Telemática
Seminario de Investigación

Evaluación y limpieza en los datos de entrenamiento: un enfoque desde la inteligencia artificial

David Camilo Corrales Muñoz
Estudiante de Doctorado
03 de octubre de 2014

1. Introducción

El propósito de la relatoría es presentar los primeros 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. Causas que generan problemas en la calidad de los datos: se dividen en procesos que traen datos externos, procesos internos que generan cambios en los datos y procesos que causan desmejoras en los datos [1], en la Figura 1 se puede observar en detalle:

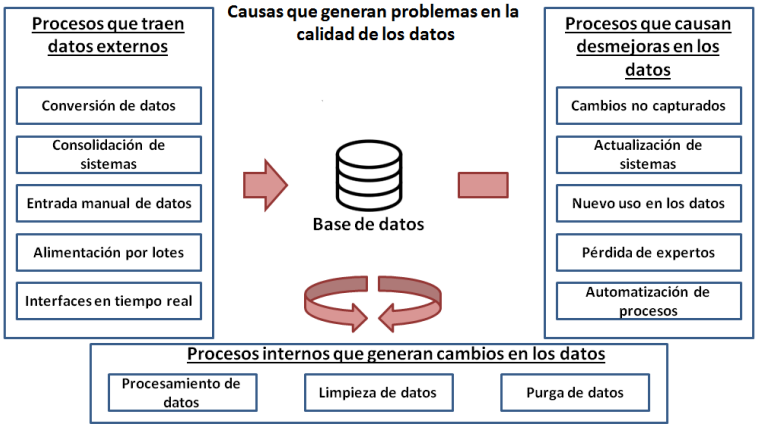

Figura 1. Causas que generan problemas en la calidad de los datos.

2.2. **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 [2]**.**

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 dos categorías (precisión y relevancia) de la taxonomía de calidad de datos en ingeniería de software empírica [3], de la siguiente manera (Figura 2):

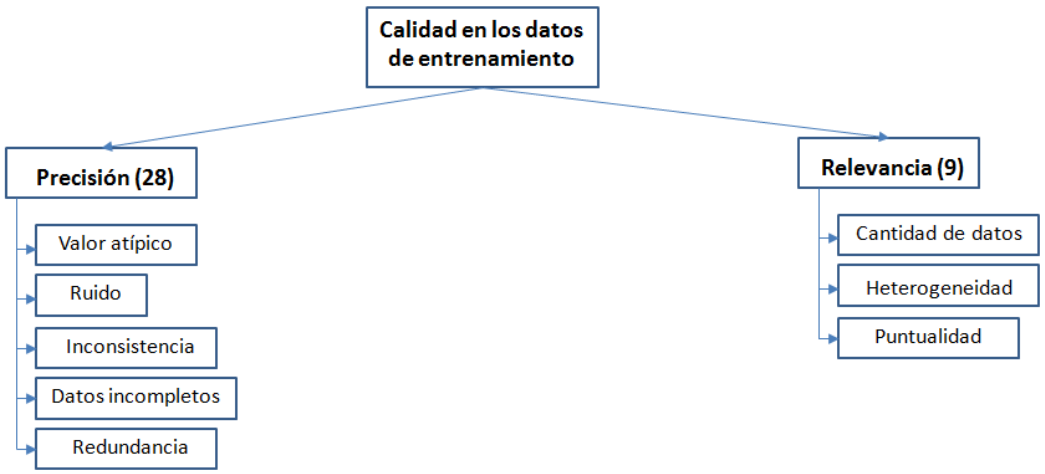

Figura 2. Taxonomía de calidad de datos en ingeniería de software empírica

De esta forma se encontraron 28 trabajos de investigación para la categoría precisión y 9 para relevancia. Los algoritmos utilizados en las investigaciones clasificadas en la categoría precisión fueron los siguientes (Figura 3):

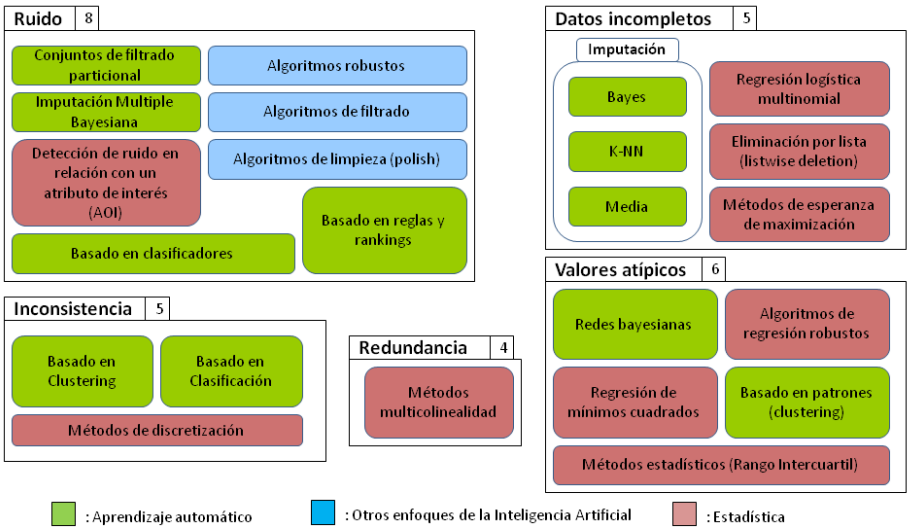

Figura 3. Algoritmos utilizados en las investigaciones clasificadas en la categoría precisión.

Los algoritmos encontrados en la Figura 3, fueron agrupados según sus características en: aprendizaje automático, otros enfoques de inteligencia artificial y estadística, llegando a la conclusión que la mayoría de artículos antiguos provienen de métodos estadísticos, las técnicas de aprendizaje automático 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 la categoría precisión.

Por otra parte los algoritmos utilizados en las investigaciones clasificadas en la categoría relevancia fueron los siguientes (Figura 4):

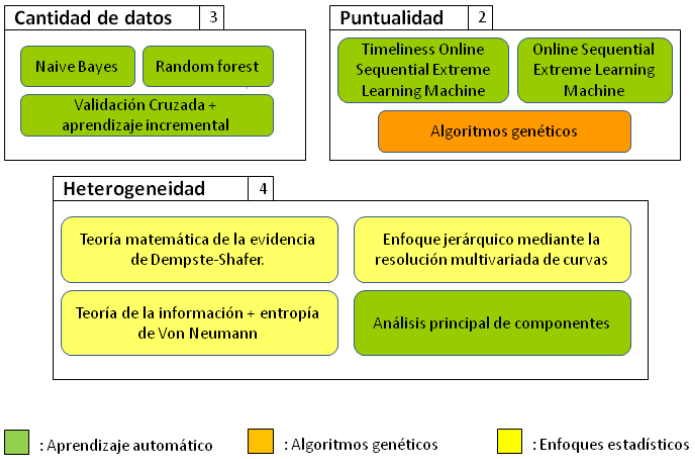

Figura 4. Algoritmos utilizados en las investigaciones clasificadas en la categoría relevancia

Los algoritmos encontrados en la Figura 4, fueron agrupados según sus características en: aprendizaje automático, algoritmos genéticos, y enfoques estadísticos, llegando a la conclusión que las técnicas de aprendizaje automático 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 la categoría relevancia.

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 marco de referencia para la evaluación y limpieza 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 precisión y relevancia en 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, según los criterios de precisión y relevancia.

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

Bibliografía

[1] A. Maydanchik, *Data Quality Assessment*: Technics Publications, LLC, 2007.

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

[3] M. F. Bosu and S. G. MacDonell, "A Taxonomy of Data Quality Challenges in Empirical Software Engineering," in *Software Engineering Conference (ASWEC), 2013 22nd Australian*, 2013, pp. 97-106.