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

Relator: David Camilo Corrales Muñoz, estudiante de Doctorado
Co-relator: Juan Carlos Corrales
Protocolante: Gustavo Andrés Uribe Gómez, estudiante de Doctorado

Fecha: 5 de Diciembre de 2014
Hora inicio: 10:11 a. m.
Hora fin: 12:03 m.

Lugar: Salón de posgrado, FIET, Universidad del Cauca, Popayán

Asistentes:
Dr. Juan Carlos Corrales, coordinador del seminario y co-relator
Ing. Camilo Corrales, estudiante de Doctorado, relator
Estudiantes de Maestría y Doctorado en Ingeniería Telemática
Estudiantes de pregrado de la FIET

Orden del día:
1- Presentación a cargo del relator
2- Intervención del co-relator
3- Discusión
4- Conclusiones

Desarrollo

1- Presentación a cargo del relator

El ingeniero Camilo Corrales, presentó el avance de su trabajo de Doctorado, para lo cual había preparado la siguiente agenda:

- Contexto
- Escenario de motivación
- Trabajos relacionados
- Objetivos

En la introducción se presentó en primer lugar el contexto de la investigación el cual es la evaluación y mejora de la calidad de los datos. En el contexto inició su presentación explicando lo que es el aprendizaje supervisado indicando que es el proceso en el cual un

algoritmo aprende a partir de conjunto de ejemplos (datos de entrenamiento), con la intensión de predecir o clasificar un nuevo dato de entrada. Los datos de entrenamiento estan compuestos de un conjunto de atributos y la clase (varible objetivo) que corresponde a tales atributos. Un conjunto de estos valores recibe el nombre de instancia. Dichos datos de entrenamiento son usados por un algoritmo que los procesa y obtiene un clasificador o modelo. Este clasificador obtine una clase a partir de unos datos de entrada suministrados. Los algoritmos de aprendizaje supervisado más utilizados son: Las máquinas de vector de soporte (SVM), redes neuronales artificiales (ANN), árboles de decisión (DT), vecino más cercano (K-NN) y redes bayesianas (BN).

A continuación se presentó el concepto de framework. Este se definió a nivel general como la representación de los componentes principales de un sistema o problema de interés, mostrando sus interrelaciones o vínculos. Sirve para desarrollar un entendimiento común para abordar un tipo de problema (Georgina et al, 2011). En el caso específico de los frameworks para la calidad de los datos se mostraron las siguientes definiciones:

- Herramienta para evaluar la calidad de los datos dentro de una organización (Wang et al, 1996).
- 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 (Willshire, et al, 1997).
- Un framework debe evaluar, y proporcionar un esquema para analizar y resolver problemas de calidad de datos para una gestión proactiva (Eppler et al, 2000).
- Buscan evaluar áreas donde los procesos de baja calidad reducen la rentabilidad de una organización (Kerr et al, 2004).

Dentro de las opciones mostrada se considera que la tercera definición se ajusta más al trabajo que se esta planteando.

Posteriormente, el relator procede con el siguiente punto de su agenda, es decir el escenario de motivación. Se indica que el aprendizaje supervisado puede ser aplicado a diversas áreas como lo son la salud, las finanzas, la predicción del clima y la agricultura de calidad. En esta sección se expone que si los datos de entrenamiento no son de buena calidad entonces el clasificador resultante no sera funcional. Los sistemas informáticos adolecen de mecanismos que evalúen la calidad de los datos. Las causas más conocidas de los problemas en los datos de entrenamiento son:

1. Entradas manuales de los datos
2. Consideración de datos de diferentes fuentes
3. Obsolecencia de los datos
4. Captura de datos en tiempo real (problemas con los sensores)

A continuación se exponen los trabajos relacionado. El primer trabajo realcionado presentado es la metodología CRISP-DM usada para la minería de datos (Chapman et al, 2000). Se presentaron brevemente las fases y tareas definidas por esta metodología. El trabajo presentado se va a centrar en las siguientes fases:

- Comprensión de los datos
- Preparación de los datos

Dentro de estas fases se tendrán en cuenta la siguientes tareas:

1. Verificar la calidad de los datos
2. Seleccionar los datos
3. Limpiar los datos
4. Construir los datos

En la primera tarea se diagnostican los problemas de los datos. La segunda fase se realiza de manera manual o asistida. Luego se realiza la limpieza de los datos que tienen problemas. Esta fase se realizara igualmente de manera asistida. En la construcción de los datos se realiza una reducción de la instancias y atributos de los datos. A continuación se presenta una taxonomía publicada por Bosu y MacDonell en el año 2013, en donde se describen las características de los datos de entrenamiento. En el primer nivel de la taxonomía encontramos la precisión, relevancia y procedencia de los datos. Estas características son las que se desea mejorar mediante procesos de limpieza de datos, sin embargo la procedencia de los datos se considera una característica fuera del ámbito de la propuesta del relator debido a que es un aspecto que le concierne al domino comercial (ventaja competitiva de los datos). Bajo el concepto de precisión encontramos: los valores atípicos, ruido, inconsistencias, datos incompletos y redundancia. Bajo el concepto de relevancia encontramos: cantidad de datos, heterogeneidad y puntualidad (información reciente). Con base en la metodología y la taxonomía mencionada se presentaron los demas trabajos relacionados.

El relator presentó 38 trabajos relacionados con la pérdida de datos, 34 de valores atípicos, 18 relacionados con ruido y 14 con la inconsitencia de los datos. De los 38 trabajos relacionados con la pérdida de los datos se extrajeron las técnicas usadas para la verificación de la falta de datos y para la limpieza de los datos faltantes. Las técnicas se clasificaron de acuerdo a las siguientes clases: aprendizaje supervisado, aprendizaje no supervisado, métodos de imputación y otros (ontologías). De igual manera los 34 trabajos relacionados con los valores atípicos se extrajeron técnicas para la detección de estos valores y se clasificaron acorde a las siguiente clases: métodos estadísticos, aprendizaje supervisado, aprendizaje no supervisado y otros. De los 18 trabajos relacionados con el ruido se extrajeron técnicas para su detección y para la detección de instancias duplicadas. Estas técnicas se clasificaron en las siguientes categorías: aprendizaje supervisado, aprendizaje no supervisado y otros. De los 14 trabajos relacionados con la inconsistencia de los datos se extrajeron técnicas que se clasificaron en las siguientes categorías: aprendizaje supervisado y métodos estadísticos.

Después de presentar estos trabajos relacionados, el relator presentó su pregunta de investigación, la cual esta redactada como: ¿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?.

A continuación, se siguió con la siguiente sección de la agenda, es decir los objetivos. Estos se presentaron así:

- 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:
    1. Establecer mecanismos para la evaluación de la calidad de los datos de entrenamiento, mediante de técnicas de Inteligencia Artificial.
    2. Definir técnicas de limpieza en los datos de entrenamiento basado en algoritmos de Inteligencia Artificial.
    3. Seleccionar estrategias para reducir la dimensionalidad en los datos de entrenamiento a partir de algoritmos de Inteligencia Artificial.
    4. Desarrollar y evaluar experimentalmente un prototipo que valide la aproximación propuesta.

Con la presentación de estos objetivos terminó la presentación del relator.

2- Intervención del co-relator

El doctor Juan Carlos Corrales, amplía información del estado actual del proyecto. El correlator presenta la dificultad de definir el concepto de framework debido a su uso en diversos ámbitos y su difícil traducción a nuestro lenguaje. Se trabajo con la definición de framework conceptual, pero se debe definir mejor los términos que se van a usar para el concepto de framework. Por otro lado, se indicó que no se piensa descartar la relevancia de los datos para así incluir la cantidad de datos como un factor a evaluar. Los objetivos de la propuesta y la metodología ya han sido fijadas.

3- Discusión

El estudiante de doctorado Gustavo Uribe pregunta, ¿Que riesgos se corren al ignorar las fases y tareas previas pertenecientes a la metodología CRISP-DM?. El relator responde que las fases y tareas previas tienen como objetivo brindar un conjunto de datos con atributos relevantes para el negocio. El proyecto actual dará por hecho que los datos ya cumplen esta característica, dado que estas fases previas corresponden a un trabajo ínter-disciplinar fuera del alcance del proyecto.

El estudiante de doctorado Diego Durán pregunta que si ha considera que ocurran fallas dentro de las fases previas. El relator responde que esto se tiene en cuenta dentro de las fases mencionadas y abordadas por la propuesta. Una tarea relevante para esto es la integración de los datos, la cual esta considerada en la metodología.

El estudiante de doctorado Mario Solarte sugiere que no se deje por fuera la base de datos bibliográficos de Google Scholar ni la de ACM, pues estas cuentan con trabajos muy recientes que pueden ser de gran relevancia para el trabajo. Adicionalmente el estudiante de

doctorado realiza la pregunta, ¿Porqué no hablan de datos de entrenamiento en esta nueva propuesta?. El relator responde que debe incluirse para aclarar a que tipos datos se refiere.

El estudiante de doctorado Gustavo Uribe pregunta, ¿Como puede influir el hecho de que se usen técnicas de aprendizaje supervisado para la limpieza de los datos de entrenamiento de otro sistema de aprendizaje supervisado? ¿Esto no desencadena un circulo vicioso?. El relator dice que podría ser, pero se pueden complementar con otro tipo de metodologías para evitar esta problemática. El co-relator aporta que es completamente factible y común usar los mismos algoritmos para la limpieza y luego para la clasificación.

4- Conclusiones

El coordinador del seminario da fin a la discusión y procede a emitir las conclusiones:

El trabajo esta avanzado en su proceso de definición y se esta realizando el levantamiento del estado del arte, ya se han definido los objetivos del trabajo pero falta pulir algunos aspectos de la propuesta .

Se termina la sesión.