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

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

Fecha: 3 de Octubre de 2014
Hora inicio: 10:07 a. m.
Hora fin: 11:05 a. 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:

- Introducción
- Motivación
- 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 se indicó las principales causas de los problemas en la calidad de los datos para una base de datos. En el contexto de

la conversión de datos, es decir la conversión de textos planos y hojas de cálculo a una base de datos, se mostró que la falta de meta-datos deriva en baja calidad de los datos. En el contexto de la consolidación (integración) de sistemas se mostró que frecuentemente los datos son duplicados, se solapan o no se ajustan. En el contexto de la entrada manual de datos, la entrada se hace a través de formularios y es muy frecuente que los datos introducidos incluyan errores de diversos tipos. En el contexto de la alimentación por lotes, se transfieren bloques de datos de manera automática y usualmente proveniente de diversas fuentes. Dado que no se tiene mayor control en este proceso se originan diversos problemas en los datos. En el contexto de envío de datos en tiempo real, los dispositivos envían constantemente información y una falla puede generar múltiples datos erróneos. En el contexto del procesamiento de datos para generar información, se mostró que una falla en los algoritmos puede causa múltiples errores en la información. En el contexto de la limpieza de datos, se mostró que los algoritmos actuales para esta tarea no garantizan conservar la integridad de todos los datos. En el contexto de la purga de datos, donde se pretende borran datos obsoletos, accidentalmente se puede eliminar datos relevantes.

Por otro lado se presentaron algunos procesos que causas desmejoras en los datos. El primero de estos procesos son los cambios no capturados. Es decir, la no actualización de los datos, por ejemplo después de una mudanza. El segundo proceso es la actualización de sistemas; en este caso los meta-datos mal elaborados ocasionan que las actualización automáticas deriven en mal construcción de los datos. El tercer proceso es el nuevo uso de los datos, en este caso se debe tener en cuenta que la imprecisión de los datos en un contexto puede ser tolerable mientras que en otro no. El cuarto proceso es la pérdida de expertos, en donde hay fuga de conocimiento, el cual puede ser imprescindible para la correcta interpretación de los datos. El quinto proceso es la automatización de procesos y básicamente incluye los procesos de automatización anteriormente mencionados.

Pese a que estos problemas se han identificado en las bases de datos, estos son aplicables para los datos de entrenamiento que son el objeto de estudio en la propuesta de trabajo presentada. Se explica que los datos de entrenamiento describen un conjunto de atributos y una variable objetivo que se quiere predecir o clasificar. Estos datos se usan como entrada para un algoritmo de aprendizaje supervisado, el cual generará un clasificador o hipótesis. 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).

Posteriormente, el relator procede con el siguiente punto de su agenda, es decir la motivación. 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. A continuación se expone, 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 taxonomía mencionada se presentaron los trabajos relacionados.

El relator presentó 28 trabajos relacionados con la precisión de los datos de entrenamiento y 9 concernientes a la relevancia. Cada uno de los trabajos usa diferentes aproximaciones para realizar el limpiado de los datos. El estudiante de doctorado identificó tres enfoques en estas aproximaciones, las cuales son: aprendizaje automático, otros enfoques de la inteligencia artificial y la estadística. De los trabajos relacionados con la precisión se concluyó lo siguiente:

1. Las contribuciones de años anteriores al 2004 tienen un enfoque estadístico
2. El enfoque de aprendizaje automático se ha mantenido desde hace un par de años.
3. Los trabajos con otros enfoques de la inteligencia artificial son una nueva tendencia.
4. No existen trabajos que cubran todos los problemas derivados de la precisión.

Algunos trabajos concernientes a la relevancia de los datos usan algoritmos genéticos, los cuales tienen un alto rendimiento. De los trabajos relacionados con la relevancia se concluyó:

1. El enfoque de aprendizaje automático se viene trabajando recientemente hasta la actualidad
2. Los enfoques matemáticas son aún relevantes
3. Los algoritmos genéticos permiten solucionar problemas de puntualidad en los datos mediante optimización
4. No existen trabajos que cubran todos los problemas derivados de la relevancia.

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 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:
    1. Establecer mecanismos para la evacuación de la precisión y relevancia en la calidad de los datos de entrenamiento, mediante técnicas de inteligencia artificial.
    2. Definir técnicas de limpieza en los datos de entrenamiento basado en algoritmos de inteligencia artificial, según los criterios de precisión y relevancia.

3. 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 proyecto, indicando que es un proyecto en colaboración con la Universidad Carlos III. El co-relator también indica que desde los años 90 se ha trabajado en la calidad de los datos en las bases de datos, sin embargo no se ha consolidado un marco de referencia para la calidad de los datos de entrenamiento. Se ha identificado esta brecha, pero aún esta pendiente delimitar los aportes que deberán incluirse en la propuesta del estudiante de doctorado (relator).

3- Discusión

El estudiante de doctorado Helder Castrillón pregunta, ¿Como se va presentar el marco de referencia? y ¿Cómo se va a tener en cuenta los estándares de calidad de datos?. El estudiante de doctorado (relator), responde que existen estándares de calidad de los datos, pero hasta ahora no se ha encontrado nada relacionado con la calidad de datos de entrenamiento. Se ha encontrado recientemente taxonomías y ontologías que pueden ser usadas como estándares para validar los resultados obtenidos. El doctor Juan Carlos Corrales, responde a la primera pregunta indicando que actualmente se ha hablado de marco de referencia, sin embargo, se podría usar el termino en inglés “framework” o el término mecanismo. Al final lo que se quiere obtener es un conjunto de técnicas asistidas que permitan evaluar la calidad de los datos de entrenamiento.

A continuación, el estudiante de doctorado Diego Duran realiza la pregunta, ¿Porqué la propuesta no tiene un contexto definido, siendo que se desea realizar una aplicación en la agricultura?. El relator responde que la técnicas usadas son independientes de contexto y se va ha aplicar en diversos dominio donde sea factible obtener un buen volumen de datos. El contexto de aplicación de la agricultura estará incluido, pero dada la dificultad de obtener datos en este dominio, esto no será un limitante.

El estudiante de doctorado Gustavo Uribe realiza la pregunta, ¿El enfoque de tu trabajo va ha ser netamente estadístico o se va ha incluir otros mecanismos como las ontologías?. El relator responde que la clasificación de enfoques presentada tenía un propósito de clasificar los trabajos relacionados de acuerdo a ciertos temas, sin indicar los métodos que finalmente se van a usar. En cuanto al uso de ontologías se ha discutido pero estas se asocian a un dominio en particular, lo cual no es deseable.

Adicionalmente, el estudiante de doctorado Gustavo Uribe solicita ampliar la explicación de los problemas causado por los datos generados en tiempo real y de la automatización de procesos. El relator y el co-relator dan una serie de ejemplos en los que se presentan medidas de tiempo real o de procesos automáticos que generan datos erróneos.

Por último, el estudiante de doctorado Helder Castrillón pregunta: ¿Porque la solución es la inteligencia artificial?. El relator responde que pueden existir diversas aproximaciones para solucionar el problema pero se tiene un interés científico en la aproximación presentada.

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, sin embargo aún falta delimitar mejor los aportes del trabajo.

Se termina la sesión.