**Universidad del Cauca**

**Facultad de Ingeniería Electrónica y Telecomunicaciones**

**Programas de Maestría y Doctorado en Ingeniería Telemática**

**Seminario de Investigación**

**MULTICLASIFICADOR PARA LA DETECCIÓN DE ROYA EN CULTIVOS DE CAFÉ EN COLOMBIA**

**Relator: David Camilo Corrales Muñoz, estudiante de Maestría**

**Co-relator: PhD. Ing. Juan Carlos Corrales Muñoz**

**Protocolante: Daniel Alberto Jaramillo Morillo, estudiante de Maestría**

**Fecha:** 07 de Marzo de 2014

**Hora inicio:** 10:10 a. m.

**Hora fin:** 11:20 a. m.

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

**Asistentes:**

Ph.D. Juan Carlos Corrales M., coordinador del seminario

Ing. David Camilo Corrales Muñoz, estudiante de Maestría, relator

Estudiantes de Doctorado y Maestría en Ingeniería Telemática

**Orden del día:**

**Presentación del seminario a cargo de Juan Carlos**

1. Presentación a cargo del relator
2. Intervención del co-relator
3. Discusión

**Desarrollo**

1. **Presentación a cargo del relator**

El ingeniero David Camilo Corrales Muñoz, presentó su tesis de Maestría, para lo cual había preparado la siguiente agenda:

* Introducción
* Motivación
* Objetivos
* Propuesta
* Evaluación
* Contribuciones, Conclusiones y Trabajos Futuros

En la Introducción, el relator realizó una breve explicación del concepto de aprendizaje supervisado, proceso en el cual un algoritmo aprende a partir de conjunto de ejemplos llamados datos de entrenamiento, con la intensión de predecir o clasificar un nuevo dato de entrada. El relator explico que los datos de entrenamiento se componen de un conjunto de atributos y una variable objetivo, estos son enviados a un algoritmo y de este se obtiene un clasificador el cual permite predecir. A continuación el relator menciona y explica cada uno de los algoritmos de aprendizaje supervisado: Maquinas de vector de soporte, Redes Neuronales Artificiales, Arboles de decisión, Vecino más cercano y Redes Bayesianas.

Luego, el relator explica el término Multiclasificador como un conjunto de clasificadores, cuyas predicciones individuales se combinan de alguna manera para obtener una predicción final en conjunto, luego habla de las Arquitecturas, los Métodos de Enfoque y las estrategias de combinación utilizados por estos Multiclasificadores.

En la Motivación, el relator mencionó que el Grupo de Estudios Ambientales presento la problemática con roya en Zupracafé en específico la Hacienda los Naranjos. De la investigación en el tema se encontró que la roya es la responsable del 30% de la perdida en los cultivos de café en Colombia. El relator explica como el agricultor de manera manual detecta el riesgo, plantea las estrategias de mitigación y calcula los costos de inversión. Sin embargo, el problema se presenta por que el proceso es manual y demorado, de la misma forma las estrategias de mitigación llegan tarde y se suben los costos de inversión.

El relator presenta la curva de distribución de riesgo, que se encuentra fragmentada en Pérdidas Esperadas, Pérdidas Manejables y Pérdidas de Estrés, en donde las últimas son catastróficas y llevarían a la perdida de todo el cultivo del café.

Luego habló de los trabajos relacionados en los que se menciona artículos en donde se trata cultivos de Arroz, Mango y Café con métodos Empíricos, Híbridos y de Ensamble, en estos se encontró la carencia o redundancia de atributos en los datos de entrenamiento. Además, que los clasificadores Bagging generan poca predicción. En este sentido el relator muestra la pregunta de investigación:

¿Cómo detectar la roya en cultivos de café Colombiano haciendo uso de clasificadores, un multiclasificador idóneo y la combinación de salidas pertinentes?

Luego expone los objetivos planteados de la siguiente manera:

Objetivo general:

* Desarrollar un multiclasificador para la detección de episodios de roya en cultivos de café en Colombia.

Objetivos específicos:

* Seleccionar los clasificadores (AD, RB, RNA, SVM, K-NN) conforme a los datos de entrenamiento que describen la ocurrencia de la roya en el cultivo de café en Colombia.
* Definir el tipo de multiclasificador (Bagging, Boosting, Stacking, Cascading, Híbridos, etc.) considerando n - niveles de clasificadores base.
* Definir la estrategia de combinación de las salidas (resultados) para los clasificadores seleccionados (Métodos de nivel abstracto, rango, medias, meta clasificadores, etc.).
* Desarrollar y evaluar experimentalmente un prototipo que implemente las capacidades del multiclasificador para la detección de roya en cultivos de café en Colombia.

Después el relator presenta su propuesta, en donde da a conocer la metodología utilizada: CRISP-DM, la cual cuenta con 6 fases y un mecanismo de doble vía, pasando siempre por el refinamiento de datos. Entre las fases se tiene la compresión del Negocio, aquí se definió la tasa de infección, en la fase de compresión de los datos la información de la hacienda Los Naranjos, para la fase de preparación de los datos se utilizó Excel y Weka, en la fase de modelado el relator presenta el Multiclasificador para la detección de roya y por ultimo menciona la fase de evaluación en donde se utilizaban métricas de clasificación, métricas de regresión y validación cruzada. Además, el relator explica las 5 componentes del multiclasificador, en los cuales se tiene dos módulos y tres clasificadores, uno de primer nivel y dos de segundo nivel.

Luego el relator da a conocer los datos de entrenamiento obtenidos en la hacienda Los Naranjos, la cual tiene 37 lotes, de ellos se escogieron los lotes más cercanos a la estación meteorológica. Los datos se obtuvieron teniendo en cuenta la variable objetivo, ósea la tasa de infección de la roya y 21 atributos entre los que se tiene, condiciones climáticas, estado del suelo, estado del cultivo, control del cultivo, y se obtuvieron 147 instancia que se tomaron desde el 2011 hasta el 2013. El relator menciona que la tasa de infección se roya por lote, se calcula observando las hojas de 60 árboles, luego se saca el porcentaje dividiendo el número de hojas infectadas, sobre el número de las hojas presentes en los 60 árboles.

Después, el relator explica cómo se construye el primer módulo, el cual es el filtro para la detección de valores atípicos con base en unos valores atípicos leves y unos valores atípicos extremos. Luego explica el modulo para la detección de patrones, el cual se encarga de buscar patrones sobre un conjunto de datos, para ello se utiliza el algoritmo K-medias con una k=3 y posterior se utilizó el clasificador bayesiano para ver la distribución de datos sobre esos datos de entrenamiento, para saber cómo funcionaban esos patrones. Con esto se obtuvieron resultados para tres grupos, a partir del análisis de estos, se elige el clasificador de primer nivel, donde se explica que este elige cuál de los dos clasificadores de segundo nivel predecirá la tasa de infección de roya, luego se elige el primer clasificador de segundo nivel, clasificador experto de predecir cuando ocurra una posible amenaza de roya y para finalizar el relator nombra el segundo clasificador de segundo nivel, este se centra en predecir cuándo hay una baja amenaza de roya. Y con esto el relator da por cumplidos los dos primeros objetivos específicos.

Luego el relator procede a explicar el clasificador de primer nivel, aquí hace referencia a la utilización de un algoritmo de propagación hacia atrás, en donde se recibe un porcentaje del error basado en la contribución que haya aportado cada neurona a la salida original, en pocas palabras introduce una instancia a la red neuronal y esta se propaga desde la primera capa hasta lograr una salida y se compara con el valor deseado para poder obtener un margen de error y distribuirlo a las capas de atrás con el fin de calibrar el clasificador.

A continuación el relator describe el primer clasificador de segundo nivel, el cual hace referencia al árbol de regresión M5, este divide el conjunto de datos en zonas mediante los valores que toman los atributos. Luego explica la diferencia de un árbol tradicional a un árbol de regresión lineal, pues precisamente este genera un modelo de regresión lineal para cada hoja, y posteriormente a través del modelo de regresión generado, calcula el valor de la predicción.

Después, explica el segundo clasificador de segundo nivel, en donde se habla de las máquinas de vector de soporte para regresión, lo que se hace es encontrar la distancia adecuada entre los vectores de soporte y el hiperplano óptimo de decisión, que reúna el mayor número de elementos. Por otra parte los elementos que están por fuera de los vectores de soporte, a una distancia dada, no son tenidos en cuenta debido a que añaden ruido. Y con esto el relator da por cumplido el objetivo específico número tres.

Aquí, el coordinador del seminario le indica al relator que le quedan 5 minutos para terminar, entonces el relator a partir de aquí comienza a mostrar de una forma más rápida su presentación y habla acerca de la evaluación, en donde explica que se utilizó una matriz de confusión para tres clases, Caturra, Castillo y Bourmon. Luego, explica las métricas de evaluación en donde se tiene la precisión o capacidad del clasificador para evitar el ruido, exhaustividad o Instancias relevantes clasificadas y medida-F o balance entre la precisión y exhaustividad.

El relator presenta con rapidez el espacio ROC, las curvas de ROC, el coeficiente de colaboración de Pearson y a continuación explica la validación cruzada, como una técnica para evaluar los resultados del clasificador, dividiendo el conjunto de datos de entrenamiento en k subconjuntos y así obtener los resultados de la evaluación del clasificador de primer nivel, con resultados de precisión mayor al 88.5%, exhaustividad mayor al 87.4%, después se evaluó el primer clasificar de segundo nivel, desafortunadamente los errores fueron muy altos, no era lo esperado pero debido al conjunto mínimo de instancias, ocurre este comportamiento. Lo mismo se hizo para el segundo clasificador de segundo nivel, también se obtienen errores de predicción muy altos y se escoge la máquina de vector de soporte para este clasificador, ya que fue la que presento menos problemas, y con esto se dieron por alcanzados todos los objetivos.

Luego, el relator menciona las siguientes contribuciones:

* Un conjunto de datos de entrenamiento que involucra información climática, estado del suelo, estado del cultivo, y control del cultivo, recolectados en la granja experimental los Naranjos.
* Un conjunto de clasificadores base idóneos para la detección de la incidencia de la roya en el café.
* Un multiclasificador con una arquitectura hibrida, conformado por dos niveles y tres clasificadores distribuidos de la siguiente forma: un clasificador en secuencia y dos en paralelo con el propósito de detectar episodios de roya en cultivos de café en Colombia.

En la parte de Conclusiones el relator menciona, que para la construcción de un multiclasificador es de vital importancia conocer el comportamiento que tienen los datos de entrenamiento, también menciona que el número bajo de instancias para entrenar cada uno de los clasificadores base, limitan el desempeño del multiclasificador, no se puede tomar la decisión más acertada si no cuenta con los datos necesarios, otra conclusión es que las causas que propiciaron altos errores de precisión en la predicción de la tasa de infección de roya fueron los datos que vinieron de diferentes fuentes, algunos empíricos y otros por la estación meteorológica. Y para finalizar con las conclusiones, la metodología para el cálculo de la tasa no es la adecuada, ya que en la hacienda los Naranjos cada lote tiene como máximo 6000 plantas y como mínimo 3000, sin embargo la metodología dice que se deben evaluar solo 60 plantas, entonces la muestra no muestra toda la distribución de datos.

Como trabajos futuros el relator propone, plantear una metodología para el cálculo de la tasa de infección de roya, automatizar el proceso de recolección de datos, proponer un marco para la evaluación de datos y por ultimo construir un multiclasificador para la detección de roya basado en cuantificadores.

1. **Intervención del co-relator**

El co-relator menciona que el aumento de precio del café ha evolucionado primero por la sequía que se ha tenido en Brasil y segundo por todas las pérdidas que se han tenido en Centro América debido a la roya, lo otro es que Colombia se está conectando con las especialidades y los mejores sabores y olores son de variedades susceptibles a la roya.

1. **Discusión**

El Ingeniero Iván Hernández resalta que las diapositivas estaban muy bien hechas, sim embargo, se tenía que mejorar en el manejo del tiempo, también menciona que el relator todo el tiempo se enfocaba en la hacienda los naranjos y de un momento a otro pasa a Colombia. A esto el relator aclara que en el escenario de motivación se enfocan los porcentajes de roya para Colombia pero que los datos de entrenamiento fueron recogidos en la hacienda los Naranjos. Por ultimo pregunta que por que no hubo la implantación, a lo que el relator responde, que debido al bajo número de instancias se tenían datos muy pobres y de esta manera no se atrevía a implementar el multiclasificador en la finca.

El Doctor Álvaro Rendo pregunta: ¿cuándo se hace la recolecta de datos de forma manual que tipo de información manejan los granjeros y esto se podría hacer de una manera automatizada? A lo que el relator responde que al no contar en la universidad con experto que maneja este tema, pensó que la información recolectada era la adecuada, ya que la recolección de estos datos es un proceso muy complicado por lo extenso de la hacienda, ya se percató del problema durante la pasantía en España.

El Magíster Mario Solarte, preguntó que si ¿se tenía conciencia de la falta de datos, se sabía de la posible situación o salió de golpe durante el desarrollo del trabajo? Se identificaron inicialmente que había trabajos relacionados en donde se utilizaban estas técnicas con datos bajos o pocas instancias, entonces se tomó esos trabajos de punto de partida, pero ya en la realización de la pasantía en la Universidad Carlos Tercero de Madrid ya se percató de la ausencia de datos.

El Doctor Gustavo Ramírez preguntó: ¿De alguna manera como ejercicio para intentar cubrir que no se tuvieran tantos datos, se usó alguna forma para conseguir estos datos y así utilizar el algoritmo para verificarlo?, el relator responde que lo importante son los datos de entrenamiento, inicialmente se vieron esos atributos y con herramientas estadísticas cómo se comportan esos datos.

Luego, de esto se termina la sesión.