**Universidad del Cauca**

**Instituto de postgrados en Ingeniería Electrónica y Telecomunicaciones**

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

**Seminario de Investigación**

***Titulo de la relatoría:***

**Towards a Selection Mechanism of Relevant Features for Automatic Epileptic Seizures Detection**

**Relator: M.Sc. Maritza Fernanda Mera Gaona, estudiante de Doctorado**

**Co-relator: Ph.D. Diego Mauricio López**

**Protocolante: M.Sc. Carlos Felipe Estrada Solano, estudiante de Doctorado**

**Fecha:** 06 de mayo de 2016

**Hora de Inicio:** 10:00 a.m.

**Hora de Finalización:** 11:00 a.m.

**Lugar:** Universidad del Cauca, Popayán

**Asistentes:**

Ph.D. Juan Carlos Corrales, coordinador del seminario

Ph.D. José Luis Arciniegas, co-relator

M.Sc. Maritza Fernanda Mera Gaona, relatora

Estudiantes de Maestría y Doctorado en Telemática, Universidad del Cauca

Estudiantes de Pregrado de la FIET, Universidad del Cauca

**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:**

M.Sc. Maritza Mera inicia la presentación mencionando que su artículo será presentado en agosto en una conferencia internacional en Munich. Luego, nombra a su director, el Ph.D. Diego López, y a su asesor, el Ing. Rubiel Vargas.

Como agenda de la presentación expone los siguientes elementos: la introducción, el planteamiento del problema, el análisis de relevancia, el mecanismo desarrollado, la discusión de los resultados y las conclusiones.

**Introducción**

La M.Sc. Mera explica que las señales de electroencefalograma (EEG) son un método electro-fisiológico de medida de actividad cerebral, en donde se utiliza un dispositivo que captura las corrientes de naturaleza iónica presentes en la corteza cerebral a partir de unos electrodos colocados en el exterior del cráneo. La señal eléctrica recogida, se amplifica y se representa en formas de ondas a lo largo del tiempo, sobre las cuales se busca encontrar anomalías epileptiformes.

La relatora aclara que las anomalías epileptiformes no necesariamente son convulsiones epilépticas o eventos que ocurren a personas epilépticas. En la literatura se encuentra que el 5% de la población puede tener un EEG anormal sin padecer ningún tipo de epilepsia. De esta forma, define las anomalías epileptiformes como descargas agudas de mayor voltaje y rapidez que la actividad cerebral de fondo, las cuales son diagnosticadas a través de electroencefalogramas y que, junto con los datos clínicos del paciente, permiten caracterizar y clasificar el síndrome. En esta última se presenta una de las limitaciones: la necesidad de mayor información para determinar si una persona padece de epilepsia. Otra de las limitaciones es la gran cantidad de tiempo necesario para poder hacer una lectura del EEG.

**Planteamiento del Problema**

La expositora expone que las anteriores limitaciones causan problemas a nivel de diagnóstico. Primero, la dificultad de acceso a especialistas, puesto que se requiere un gran esfuerzo y mucho tiempo para la lectura de un EEG. Particularmente, en países en vía de desarrollo, como Colombia, hay un neurólogo por cada 200 mil habitantes, lo cual dificulta el acceso a este tipo de médicos para personas que viven en sitios rurales alejados. Lo anterior genera altos costos del diagnóstico para el análisis de encefalogramas.

La M.Sc. Mera expresa que desde hace algunos años se ha estado trabajando en la detección automática de convulsiones epilépticas en una señal EEG utilizando un esquema como se describe a continuación: un dispositivo captura la señal a partir de los electrodos ubicados en el cráneo de la persona, aplicando algunos algoritmos de procesamiento para eliminar el ruido y mejorar la calidad de la señal; luego, existe un proceso de extracción de características que utiliza operaciones matemáticas para analizar la señal, para finalmente realizar la clasificación. No obstante, la gran cantidad de características podría afectar la calidad del clasificador, razón por la cual propone realizar un análisis de relevancia para optimizar el esquema presentado.

Con base en lo anterior, la relatora presenta las siguientes preguntas:

* ¿Cuáles son las características que describen la señal?
* ¿Cuáles son las características relevantes?
* ¿Cómo determinar si una señal es anormal o normal?

**Análisis de relevancia**

La expositora define el análisis de relevancia como un proceso de selección de atributos para escoger un subconjunto de atributos del conjunto original. En el conjunto original se puede encontrar datos relevantes que pueden servir para la clasificación, datos irrelevantes que representan ruido para la clasificación, y datos redundantes que es necesario eliminar. El objetivo de hacer un análisis de relevancia es tener menos datos (algoritmos aprenden más rápido), mayor exactitud (el clasificador generaliza mejor) y menos atributos (menor procesamiento). Este proceso pretende elegir atributos que sean relevantes para una aplicación y lograr el máximo rendimiento con el mínimo esfuerzo.

La M.Sc. Mera explica dos enfoques para llevar a cabo el análisis de relevancia. Primero, la transformación de atributos, el cual crea un nuevo conjunto de atributos. Segundo, la selección de atributos, el cual extrae un conjunto de atributos a partir de los originales. Para este último, es importante tener en cuenta que a mayor dimensionalidad en un gran conjunto de características, menor la eficiencia del clasificador.

**Mecanismo**

La relatora menciona que para definir su mecanismo, se realizó una revisión de la literatura respecto a los tipos de bases de datos utilizadas, los principales tipos de extractores de características, la selección de características y la clasificación. En este punto, aclara que dicha revisión ya fue presentada en sesiones previas del seminario de investigación.

La expositora comenta que seleccionaron una base de datos de EEGs del Hospital Infantil de Boston, la cual, aunque es muy pequeña, es de las más utilizadas por los trabajos encontrados en la literatura. Esta base de datos contiene información de 23 pacientes (22 niños y 1 adulto), maneja un sistema 10-20 (sistema utilizado en un ámbito clínico real), provee una tasa de muestreo de 256 por segundo, utiliza el formato EDF (formato estándar para EEG) y es multicanal (23 canales).

Luego, la M.Sc. Mera comenta que por medio de la revisión de la literatura ya se había encontrado las características más comunes que se calculaban: entropía, máxima amplitud, mínima amplitud, varianza media, máxima potencia, mínima potencia y potencia media. De esta forma, con 23 canales y 7 extractores se logra un conjunto de datos con 161 características, en donde es necesario aplicar un análisis de relevancia para determinar los valores significativos para la clasificación.

Seguidamente, la relatora expone que utilizaron 4 algoritmos para clasificación y 3 algoritmos para análisis de relevancia. Los algoritmos de clasificación son: Naive Bayes, Rule Induction, KNN y Decision Tree. Los algoritmos de análisis de relevancia son: Forward Selection, Optimize Selection y Backward Elimination. De esta forma, presenta una tabla con los mejores resultados para cada uno de los algoritmos de clasificación junto con cada uno de los algoritmos de análisis de relevancia.

Con base en lo anterior, menciona que se revisó en detalle las características que se utilizaron para cada uno de los algoritmos que se tuvieron en cuenta, resaltando que las características con mejor comportamiento fue para el algoritmo Decision Tree. La expositora aclara que se utilizó un mecanismo de codificación para definir las características del conjunto de datos, en donde las de mejor comportamiento (aproximadamente 80% de precisión en la clasificación) fueron: a6 (potencia media primer canal), a27 (potencia media segundo canal), a94 (mínima amplitud del canal 14) y a121 (máxima amplitud del canal 18).

**Discusión**

La M.Sc. Mera menciona que se revisó en la literatura para comparar los resultados obtenidos. En los antecedentes, varios artículos incluyen la clasificación y el análisis de relevancia, cuyos resultados tienen una sensibilidad o especificidad alta, mucho mayor al 80% obtenida; no obstante, dichos estudios trabajaron con bases de datos experimentales (usualmente, monocanal). A diferencia de estos trabajos, la presente propuesta utiliza un menor número de características y trabaja con una base de datos real.

**Conclusiones**

Finalmente, la relatora concluye que:

* El mecanismo construido permitió obtener un conjunto de 4 características para clasificar señales EEG con un 80,77% de precisión.
* El algoritmo de selección de características empleado permitió reducir el tamaño del conjunto de características en un 97,51%.
* Los arboles de decisión arrojaron mejores resultados de clasificación con las características seleccionadas respecto a los algoritmos Naive Bayes y KNN.

**2- Intervención del co-relator**

El Ph.D. Diego López comenta que previamente se había discutido el contexto del trabajo, es decir, el enfoque de la propuesta desde el punto de vista de su aplicación. Adiciona que tenían dudas en si abordar un tema en concreto, como la epilepsia, o si realizar algo más general sin tratar específicamente un problema neurológico. Debido a lo anterior, resalta que se adaptó la propuesta para trabajar con señales epileptiformes con el fin de no enfocarse en la epilepsia, o algún otro problema en concreto, sino en identificar anomalías en una señal EEG. En este sentido, aclara que en el presente año fue aceptado un proyecto de Colciencias dentro del cual se puede desarrollar esta propuesta, lo cual obligó a centrarla en el problema de la epilepsia.

Seguidamente, el co-relator explica que se han encontrado con el problema de contar con un conjunto de datos, puesto que los necesitas para poder empezar a desarrollar la propuesta. Por lo anterior, la idea del artículo presentado es tener una aproximación al desarrollo de la propuesta de doctorado. De esta forma, les fue posible identificar una brecha en el tema de análisis de relevancia, puesto que en la literatura se encuentran muchos trabajos que utilizan clasificación para la detección automática de epilepsia, sin embargo, el análisis de relevancia propuesto en otras áreas no se ha utilizado en este campo.

Luego, el Ph.D. López menciona que el objetivo del proyecto de Colciencias es proveer un servicio de tele-encefalografía para lograr un diagnóstico remoto de señales EEG capturadas a través de dispositivos móviles en zonas rurales, permitiendo que un especialista pueda confirmar casos de epilepsia. De esta forma, comenta que se va utilizar una plataforma donde las señales EEG y los datos clínicos soportarán el diagnóstico, resaltando que van a contar con más de 161 características que pueden ayudar a confirmar los casos de epilepsia, tales como edad, condiciones fisiológicas, alimentación, etc.

Por último, el co-relator expone que aunque el conjunto de datos utilizado en este artículo fue publicado por un grupo biomédico del MIT, es muy pequeño y no tiene todas las características necesarias para un desarrollo apropiado de la propuesta. Debido a lo anterior, comenta que en el proyecto se van a recoger datos de 200 pacientes, 100 para entrenamiento y 100 para la evaluación, logrando un conjunto de datos más grande que el utilizado en el presente artículo.

Por otra parte, el Ing. Rubiel Vargas, asesor y co-autor del artículo expuesto, revela que al inicio la M.Sc. Mera estaba preocupada por los resultados obtenidos, puesto que veía valores de 95% de precisión en otros trabajos. En este sentido, resalta que los dichos trabajos utilizan únicamente un canal, mientras que en el presente artículo se utilizaron todos los canales y los descriptores más citados en la literatura. Luego, añade que la relatora tendrá que explorar otros descriptores y contar con más instancias en la base de datos, puesto que de 23 casos es muy difícil que un algoritmo aprenda; con un mayor número de instancias de aprendizaje se podrá mejorar la evaluación de la propuesta, reduciendo las características de los otros descriptores a analizar.

Por último, el asesor resalta que la base de datos utilizada en el artículo ya cuenta con un pre-procesamiento de los datos. Sin embargo, en el proyecto se espera tener los dos tipos de datos: los datos crudos y los datos pre-procesados. De esta forma, la estudiante podrá comparar la diferencia en los resultados, ya que a veces en el filtrado hardware se pierde la información que puede ser relevante para este caso.

**3- Discusión**

El Ph.D. Oscar Caicedo resalta que el artículo está bastante interesante y pregunta cuál fue el mayor aporte?

La M.Sc. Mera responde que fue la identificación de las 4 características relevantes que se obtuvieron para una precisión del 80%. Aclara que al principio estaba preocupada porque incluso encontró un trabajo que publicaba un valor de 100% de precisión. Sin embargo, explica que dicho trabajo solamente trataba la sensibilidad y que, al seguir las referencias, la especificidad no era buena, lo cual representaba una desventaja en el trabajo, puesto que dicho 100% de precisión no era tan real. Adicionalmente, menciona que el trabajo tampoco utilizaba datos reales, especificando que de haber seleccionado la base de datos empleada en el presente artículo, habrían obtenido valores más reales.

La Ing. Carolina Rico consulta si con las características identificadas se puede obviar algunas señales de algunos electrodos?

La M.Sc. Mera comenta que le faltó una diapositiva de trabajo futuro, ya que una de las hipótesis del por qué no obtuvo mejores resultados es debido a que una señal epileptiforme la clasifica como normal o anormal. No obstante, es necesario de más valores para diagnosticar la epilepsia y su foco; por ejemplo, para saber en qué posición un electrodo está señalando que hay un episodio epileptiforme (dependiendo del electrodo, el foco de la epilepsia puede variar). De esta forma, es posible identificar el tipo de epilepsia que hay en la señal. Con la base de datos actual, simplemente se informa si en un determinado segmento ocurre un episodio epiléptico, pero no el tipo. Con una mejor base de datos, se puede mejorar la clasificación, ya que no solamente se clasificará como normal o anormal, sino el tipo de epilepsia o la parte del cerebro donde se presentó el episodio anormal, permitiendo especificar más la clasificación. Al final, concluye que no se podría eliminar ninguno de los electrodos.

La M.Sc. Luz Marina Sierra cuestiona si las señales o datos que arroja el dispositivo se pueden utilizar para medir otras situaciones, como el nivel de tensión de una persona.

La M.Sc. Mera responde que realmente no sabe el dispositivo con el cual se tomaron los datos utilizados en el presente artículo, pero que seguramente es un dispositivo aprobado clínicamente que no solamente es empleado en el diagnóstico de epilepsia, sino también en el análisis de otras situaciones neurológicas.

El Ph.D. Juan Carlos Corrales pregunta por el número de instancias de la base de datos seleccionada y de la base de datos a generar.

La M.Sc. Mera responde que fueron 23 pacientes. No obstante, agrega que algunas señales EEG presentan más de un evento epileptiforme, por lo que fue posible producir más señales partir de algunos pacientes, obteniendo un total de 56 instancias.

El Ph.D. López menciona que para la nueva base de datos se pretende obtener 100 instancias para aprendizaje y 100 instancias para pruebas. La relatora agrega que, en la literatura, el trabajo con mayor número instancias era de 500; sin embargo, dicho trabajo empleaba una base de datos monocanal, lo cual no permite tener resultados reales.

El Ph.D. Corrales continúa con la discusión explicando que el tema de tener datos es el gran problema de los algoritmos supervisados, presentando dos casos no deseados: el problema de tener muchos datos, por alto costo de procesamiento, y el problema de tener muy pocos datos, por la baja capacidad de aprendizaje. Con base en lo anterior, comenta que actualmente los investigadores, desde un punto de vista de la ingeniería, están tratando de generar propuestas para disminuir el procesamiento o mejorar la precisión con pocos datos. El Ph.D. Corrales resalta que el artículo identifica unas características importantes, sin embargo, que lo hace dentro de un conjunto de datos pequeño. En este punto, aclara que se debe tener mucho cuidado, ya que usualmente con los conjuntos de datos pequeños se trata de considerar todas las características.

Continuando, el Ph.D. Corrales comenta que tiene un trabajo en el extremo de tener muchos datos (aproximadamente, 5 millones de instancias), donde el problema de crecimiento es considerable. Aclara que en su trabajo se trata de de reducir el número de instancias a utilizar para poder realizar una predicción, haciendo pruebas con el 3%, 5% y 10% del número de instancias, cada una con 41 características. De esta forma, su trabajo propone una arquitectura en serie y prueban por porcentaje del número de instancias en un conjunto de algoritmos. Por ejemplo, el primer conjunto son los árboles de decisión y en el segundo está la máquina de vector de soporte más un perceptron multicapa. Resalta que este tipo de combinaciones ya empieza a hacer mucho más importante en el trabajo. Para concluir su intervención, le recomienda a la estudiante que trate de solucionar el problema de muy pocos datos, porque incluso con el proyecto van a ser muy pocas instancias aún.

La M.Sc. Mera concuerda con que hay muchas características, pero explica que cada algoritmo no hace la clasificación con todas, sino que arranca desde un punto mínimo (una) y va aumentando. Cuando se pasa de las 15 características ya no sigue buscando porque, teniendo en cuenta la cantidad de los datos, no tiene sentido.

El Ph.D. Corrales comenta que en su trabajo pasa algo similar, puesto que inicialmente pensaron que con el 10% se darían los mejores resultados, recordando que los porcentajes varían ente 3%, 5% y 10%. No obstante, el de 3% dio mejores resultados, concluyendo que lo anterior se debe a la combinación de algoritmos y a la manera como se están combinando. De esta forma, menciona que el reto es saber qué tipo de arquitecturas se pueden juntar a partir de la ejecución de varios algoritmos por fases. Luego, recomienda enfocarse más en el problema de cómo se puede completar más instancias, incluso construyéndolas (datos sintéticos). Finalmente, agrega que, de acuerdo a la literatura, la mejor manera de construir datos sintéticos es a partir del uso de algoritmos evolutivos y genéticos.

El Ing. Vargas contesta que en el dominio clínico la relevancia de los resultados es diferente, exponiendo dos extremos: para pocos datos, se centra en la metodología, y para muchos datos, se busca la precisión. Con base en lo anterior, explica que la M.Sc. Mera propone la metodología y resalta que conseguir 200 EEGs es bastante en el campo de la clínica. Adicionalmente, menciona que construir señales EEG sintéticas es muy difícil, ya que básicamente se componen de ruido; por ejemplo, cualquier movimiento de los ojos del paciente genera señales que no son episodios epileptiformes. Por lo anterior, la M.Sc. Mera se enfocará en la parte metodológica y no en los resultados.

Por último el Ph.D. Corrales pregunta si se va a detectar el origen del evento dentro de la cabeza, puesto que es muy interesante estudiarlo a partir de la señal.

El Ing. Vargas responde que las señales utilizadas en el artículo son solamente de 20 minutos, pero que existen unas señales que ponen el dispositivo EEG al paciente para que se vaya para su casa durante 48 horas. Entonces, el especialista tiene que leer 48 horas por 23 canales. Por esta razón, se propone realizar un tipo de filtro para decirle al especialista cuáles son las señales EEG que tienen mayor posibilidad de presentar los episodios epileptiformes con el fin de que los revise primero. A partir de este punto, el especialista es quien decide si realmente se trata de un episodio epileptiforme y dónde se produjo.

**4- Conclusiones**

Las conclusiones ya fueron abordadas en la sección de discusión.

Se termina la sesión.