**Universidad del Cauca**

**Facultad de Ingeniería Electrónica y Telecomunicaciones**

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

**Seminario de Investigación**

**Detección de productos volcánicos basado en aprendizaje incremental**

**José Eduardo Gómez Daza**

Estudiante de Maestría

30 de octubre de 2015

1. **Introducción**

En esta segunda presentación en el seminario, se ha modificado ciertas partes según los avances que se han logrado en el transcurso del tiempo y las sugerencias realizadas por el público presente.

También las gráficas han cambiado ya que se han investigado más artículos y se han descartado otros.

Además, esta sección se presentó las principales definiciones de los temas del dominio de aplicación (vulcanología), sistemas dinámicos y de la parte ingenieril, como los es las definiciones de aprendizaje incremental y finalmente el aprendizaje supervisado.

1.1 Dominio de aplicación

**VOLCÁN**

Es un sistema natural que traslada energíadesde el interior de la tierra hacia el exterior, por medio de la transferencia de magma. La creación del magma ocurre en profundidad, por fusión de rocas que luego se mueven hacia la superficie hasta hacer erupción [1]

Se presentan además los productos volcánicos generados en un volcán activo, los cuales se muestran a continuación:

* Sismicidad.
* Emisión de gases volcánicos
* Caída de ceniza
* Lluvias acidas.
* Erupción.
* Flujos piroclásticos.
* Flujos de lodo.

**SISTEMA DINAMICO**

 Se considera un sistema como una entidad formada por un conjunto de elementos, que son los componentes básicos del sistema, y por las relaciones existentes entre sí y con su entorno. Cuando el sistema involucra la variable tiempo, se le denomina sistema dinámico, ya que este evolucionará a lo largo del tiempo [2].

En la figura 1 se explica lo que es un sistema, y luego se describe algunos tipos de sistemas existentes, y se presentan algunos ejemplos de cada sistema, finalmente se concluye que un sistema dinámico es a su vez un sistema estocástico y un sistema abierto que varía en el tiempo y genera datos no estacionarios.



Figura 1. Tipos de sistemas.

1.2 Informática.

**Aprendizaje supervisado**

Es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (datos de entrada y resultados deseados). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada [3].

En la figura 2 se muestra como a partir de un dataset, que contiene atributos y una clase, es posible llegar a una función objetivo pasando el dataset a un algoritmo de aprendizaje supervisado.



Figura 2. Aprendizaje supervisado.

**Aprendizaje incremental**

El aprendizaje incremental (aprendizaje online), permite analizar flujo de datos continuos (data flow) en entornos dinámicos sin necesidad de reentrenar los clasificadores. Las instancias entran al clasificador en grupos o de forma aislada y secuencial (de una en una). Descarta o reajustar continuamente, las hipótesis generadas a lo largo del proceso según aparezcan nuevos datos. Estas técnicas se caracterizan principalmente por dos aspectos: es un aprendizaje capaz de incorporar la información que aporten nuevas experiencias (que antes no estaban disponibles en el conjunto de datos) al modelo que se está induciendo y capaz de hacerlo evolucionar para que cada vez represente conceptos más complejos [6].

**Cambios de contexto**

En los sistemas dinámicos, a medida que nuevos datos van siendo adquiridos la función objetivo puede variar, de forma que reglas válidas para el pasado dejan de serlo para la actualidad. Los problemas en los cuales las causas de los cambios aparecen como ocultas son conocidos como hidden context, y sólo pueden tratarse mediante algoritmos incrementales, ya que pueden ampliar y corregir las hipótesis generadas.

El contexto oculto puede variar de forma que invalida el conocimiento ya obtenido. En estos casos, decimos que estamos en un problema con “concept drift”. El problema aquí es que al tratar con datos con probabilidad de errores o ruido, dichos datos con ruido pueden ser confundidos con un cambio en el contexto (concept drift). Los sistemas más robustos con respecto al ruido se adaptarían a los cambios con bastante retardo, pudiendo llegar a ignorar algunos. Por otro lado, un sistema muy sensible tomaría datos con ruido como cambios. Por ello, es decisivo el balance entre robustez al ruido y sensibilidad al cambio en el rendimiento de un sistema de aprendizaje con datos no estacionarios.

* **Abruptos:** Cambio repentino y significativo a la definición del proceso
* **Graduales:** Que se desarrolla o cambia en etapas sucesivas y continuas.
* **Recurrentes:** Cambio que ocurre debido a fenómenos cíclicos
* **Ruido:** Equilibrio entre sensibilidad al ruido y robustez en detección al cambio

**2. Motivación.**

Se presenta un escenario de motivación, en el cual se toma en cuenta que los volcanes son monitoreados por sensores instalados en cercanías a los volcanes, dichos sensores son pertenecientes a distintas metodologías de vigilancia, tales como geoquímica, los cuales miden concentraciones de gas radón, Co2 y So2; geofísica: los cuales son sensores que miden el movimiento de la tierra a través de sismómetros, actividad superficial: miden los cambios visibles que ocurren y son monitoreados a través de cámaras digitales y sensores climáticos.

Ahora bien, estos sensores son almacenados en repositorios de información ubicados en un observatorio vulcanológico. Si se quisiera hacer uso de las técnicas de minería de datos, específicamente de aprendizaje supervisado sobre el conjunto de datos que se tiene a través del repositorio de información volcánica, y teniendo en cuenta que los datos son adquiridos en tiempo real. Lo primero que se debería hacerse es seleccionar el conjunto de datos de un lapso de tiempo, por ejemplo de los años 2005 a 2014, luego pasar todo el conjunto de datos a un algoritmo de aprendizaje supervisado, este algoritmo nos generaría un clasificador, el cual claramente estaría entrenado con datos de los años seleccionados. Si quisiéramos entonces clasificar el tipo de ceniza que está arrojando el volcán, de acuerdo a atributos como por ejemplo tipo de sismicidad generada, duración del sismo e inflación del cono volcánico, el clasificador nos retornaría como respuesta el tipo de sismicidad que está cayendo (alta, media o baja). Sin embargo, al pasar el tiempo y como se mencionó anteriormente el volcán es un sistema dinámico que varía en el tiempo, y sus condiciones físicas cambian. El clasificador pasado este tiempo no arrojara resultados precisos, obteniendo falsos positivos en sus clasificaciones, lo cual lo convertiría en obsoleto.

**3. Trabajos relacionados.**

A diferencia de la primera presentación, la parte de trabajos relacionados en esta sección está dividida en aprendizaje supervisado con vulcanología (sismología, geodesia y geoquímica), por otra parte, el otro lado se divide en las características relevantes del aprendizaje incremental, como lo son, el algoritmo base, el tipo de adaptación al cambio, si tiene sensibilidad al ruido y finalmente si maneja ventana de instancias. Dicha representación se observa en la figura 3.



Figura 3. División para trabajos relacionados.

Los trabajos relacionados fueron consultados en los motores de búsqueda Springger, IEEE y google scholar tal y como se observa en la figura 4.



Figura 4. Motores de búsqueda consultados.

Las figuras 5. Muestra en el eje x los núcleos temáticos más importantes de la vulcanología como lo son las de simbología: clasificación de sismos, características más relevantes de las ondas y localización de sismos, además de la geoquímica y geodesia. En el eje y se encuentran el número de artículos por núcleo temático.



Figura 5. Aprendizaje supervisado en volcanología.

Sobre esta grafica 5 podemos afirmar lo siguiente:

* La sismología es el área en la cual se está trabajando más en la actualidad.
* Específicamente la clasificación de eventos sísmicos.
* Los núcleos temáticos de geoquímica y geodesia no han sido usados actualmente.

La figura 6 por su parte tiene en el eje x, los algoritmos de aprendizaje supervisado y en el eje y la cantidad de artículos. Sobre esta grafica se puede decir:

* Los algoritmos SVM y RNA han sido los más usados en el dominio de aplicación volcánica.



Figura 6. Algoritmos de aprendizaje supervisado en volcanología.

Las brechas existentes en esta sección son:

* Obsolescencia en los clasificadores.
* Conjunto de datos muy grande para reentrenar clasificadores.

Ahora bien, desde el lado del aprendizaje incremental se encuentran la gráfica 7. Aquí se encuentran los algoritmos más relevantes y están combinados los tipos de cambios agrupados por el algoritmo base del que parten los algoritmos incrementales.



Figura 7. Algoritmos de aprendizaje incremental.

Por otro lado, de acuerdo a información proporcionada por los expertos en volcanología y acoplando dicha información con aprendizaje incremental se definieron 3 áreas de vigilancia volcánica. Las áreas de vigilancia volcánica fueron clasificadas según las características de los algoritmos incrementales. De acuerdo a esto se generó la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Sismicidad** | **Geodesia** | **Geoquímica** |
| **características de algoritmos incrementales** | Clasificación de eventos | Energía liberada | Conteo de sismos | Deformación | Emisión de gases. |
| Abrupto |  | x | x | X | X |
| Gradual | X | x | x | X | X |
| Recurrente |  |  |  |  | X |
| Ruido | x |  |  | X |  |

Como conclusión en esta sección fueron seleccionados preliminarmente estos algoritmos para cada tipo de área volcánica:

**Sismología**

**Clasificación de eventos:**

Algoritmos: CVFDT, MultiCIDIM-DS,OzaBagADWIN,SEA,ACE**.**

Ninguno de estos algoritmos realiza un manejo adecuado del ruido.

**Energía liberada y Conteo de sismos:**

Algoritmos: CVFDT, IADEM.

**Geodesia:**

Algoritmos: FACIL, FLORA, FLORA2, STAGGUER

A estos algoritmos para geodesia le falta adaptarse a cambios graduales.

. **Geoquímica:**

Algoritmo: ACE

Se adapta a cambios graduales y recurrentes pero no a cambios abruptos

Se pretende tomar como punto de partida las familias de algoritmos más usadas e identificar las características más relevantes de cada algoritmo.

**4. PREGUNTA DE INVESTIGACIÓN.**

¿Cómo detectar productos volcánicos haciendo uso de algoritmos de aprendizaje incremental?

**5. HIPÓTESIS.**

El uso de técnicas de aprendizaje incremental soluciona los problemas de obsolescencia de clasificadores en los sistemas dinámicos.

**6. OBJETIVOS**

General

Construir un mecanismo que permita clasificar productos volcánicos haciendo uso de técnicas de aprendizaje incremental.

Específicos

* Definir un conjunto de datos de entrenamiento para la clasificación de productos volcánicos.
* Adaptar algoritmo(s) incremental (es) que permitan clasificar productos volcánicos.
* Desarrollar y evaluar experimentalmente un prototipo que implemente las capacidades del sistema.

**7. REFERENCIAS.**

1. V. Zobin, Introduction to volcanic seismology, Mexico, 2000.
2. A. C. Orozco Mauricio, Earthquake Research and Analysis - Seismology, Seismotectonic and Earthquake Geology, 2012.
3. Van Wyk Benjamin, evaluación y prevesion de riesgos ambientales en centro america. 1996.
4. Chouet, B.A, precursory swarm of long period events at redoubt volcano, 1989
5. Kotsiantis C, machine learning, 2007.
6. Suarez. A, Estudio e implementación en MOA de nuevos algoritmos de aprendizaje incremental basados en support vector machine, 2013.