Time series analysis of agro-meteorological through algorithms scalable data mining case: Chili river watershed, Arequipa

Abarca Romero Melisa
CLEI, 2015
Universidad Nacional de San Agustin
Arequipa, Perú
melinm2@gmail.com

Karla Fernández Fabián
CLEI, 2015
Universidad Nacional de San Agustin
Arequipa, Perú
karla.m.f.f@gmail.com

Jose Herrera Quispe
CLEI, 2015
Universidad Nacional de San Agustin
Arequipa, Perú
jherreraq@gmail.com

***Abstract*—The paper proposes a model for predicting climate change, using algorithms in mining techniques based on approximate data, applied to agro-meteorological data, by identifying groups search of motifs and time series forecasting. To achieve the goal you work with the water balance components: flow, precipitation and evaporation; also took into account the climatic variety seasons marked by humidity (December, January, February, March) and dry (other months) providing better to abstract sub-classification for temporary data processing three classification techniques: linear regression, Naive Bayes and neural networks, where the results of each algorithm are compared with other results. Then the mathematical method of linear regression predicting water balance components for a period of approximately 12 months on the data of dams Pane and Fraile Water Resources in River Basin Chili, Arequipa is performed.**

Keywords—Motifs, agro-meteorological data, flow, precipitation, evaporation, naive bayes, neural networks linear regression, and prediction.

I. INTRODUCCIÓN

Los rápidos avances en tecnologías de recolección y de almacenamiento permitieron la acumulación de grandes volúmenes de datos. Sin embargo, extraer conocimiento puede ser extremadamente desafiante. En este contexto, la minería de datos combina métodos tradicionales de análisis de datos con algoritmos sofisticados para el procesamiento de grandes volúmenes de datos [6].

La Minería de Series Temporales es un campo que incluye técnicas de Minería de Datos adaptadas para llevar en consideración la naturaleza de las series temporales en varios dominios de aplicación, como los negocios, la industria, medicina y ciencia, los procedimientos generan grandes cantidades de datos caracterizados como series temporales [8].

De acuerdo con la literatura del área, las principales tareas en minería de series temporales son: Identificación de agrupamientos, clasificación, identificación de intrusos, encontrar motifs, encontrar reglas de asociación y pronóstico. Aunque algunas de esas tareas sean semejantes a tareas correspondientes a minería de datos, el aspecto temporal coloca algunas cuestiones específicas que son consideradas y/o restricciones impuestas a las aplicaciones correspondientes. [8]

La representación, organización, y en particular, la minería de series temporales no son tareas triviales, la utilización de criterios de similitud es particularmente común en estas tareas, ya que la similitud es un concepto intuitivo para la comparación de objetos complejos. De modo específico, muchos algoritmos de minería de datos, como la clasificación, clustering, búsqueda de motifs y pronóstico dependen de algoritmos de búsqueda del vecino más próximo.

Soluciones exactas para resolver ese problema vienen siendo estudiadas ya hacía varios años, por otro lado soluciones aproximadas y probabilísticas vienen siendo poco exploradas [13].

Diversos algoritmos de minería de datos basados en la búsqueda del vecino más próximo fueron propuestos y muchos de ellos ofrecen resultados exactos. La gran cantidad de cálculos de distancia, en dominios de datos complejos, los conjuntos de objetos generalmente poseen una alta dimensión.

Las series temporales se encuentran vinculadas a variables climáticas del análisis de varios años de series temporales agro-meteorológicas a través de algoritmos escalables de minería de datos [4].

En el Marco teórico, se exponent los métodos de representación de series temporales. En la Sección III, se analizan los actividades de minería de datos se realiza el contexto hidrológico del área de estudio, para poder entender el procedimiento de la sección IV con las respectivas pruebas y resultados y finalmente las conclusiones a las que se llegaron con la investigación del tema.

978-1-4673-9143-6/15/$31.00 ©2015 IEEE

II. MARCO TEÓRICO

A. Representación de Series Temporales

Una serie temporal es una colección de observaciones hechas secuencialmente a lo largo del tiempo. En cada punto de medición en el tiempo, pueden ser monitoreados uno o más atributos, y la serie temporal resultante es llamada univariada o multivariada, respectivamente. En muchos casos, una secuencia de símbolos puede ser usada para representar una serie temporal [8].

Existen diferentes notaciones empleadas para la representación matemática de una serie temporal (1) basada en Falk y Marohn:

$$X=\{X1, X2,...\} \text{ ó } \{Xk\} \text{ donde } k >=1 \quad (1)$$

Por lo general el interés no está en las propiedades globales de la serie, sino, en las subpartes de las series, las cuales son llamadas subsecuencias.

Exite un tipo de representación tipo árbol propuesto por [10] la cual jerarquiza las series temporales. También tenemos a la Transformada discreta de Fourier (TDF), propuesta por [11] es una de las formas de representación propuesta por primera vez en el contexto de la minería de datos. DFT transforma una serie temporal a partir del dominio del tiempo para el dominio de la frecuencia. Transformada Discreta Wavelet (Discrete Wavelet Transform DWT) [13], transforma la serie temporal en espacio/frecuencia. Singular Value Descomposition (SVD) [14] realiza una transformación global, girado el eje del conjunto de datos de tal modo uqe el primer eje explica la variación máxima, el segundo eje explica el máximo de varianza restante y es ortogonal al primer eje, etc. Piecewise Aggregate Approximation (PAA) [12] divide una serie temporal en segmentos de igual longitud y registra la medida de los valores correspondientes de cada segmento [8].

B. KDD y Minería de Datos

El descubrimiento de conocimiento en base de datos (KDD), es el proceso que a partir de datos, identificar patrones válidos, nuevos, potencialmente útiles y comprensibles [8]. Según [15] este proceso es compuesto de cinco etapas: selección de datos, pre-procesamiento y limpieza de los datos, transformación de los datos, minería de datos e interpretación y validación de resultados.

Para extraer información que sea útil al usuario, es necesario un pre-procesamiento, donde la tendencia se encuentra en procesar grandes cantidades de datos, a continuación se describe estas actividades empleadas en los datos agro-meteorológicos.

C. Actividades de Minería de Series Temporales

Tareas comunes de series temporales [8], que fueron aplicadas para el caso de estudio:

- Detección de Agrupamiento: Busca grupos de series temporales en un banco de datos de modo que series temporales de un mismo grupo son semejantes unas a las otras, mientras series temporales de grupos distintos son diferentes entre sí.
- Clasificación: Atribuir una serie temporal a una clase pre-definida de modo que la serie sea más parecida con las series de esa clase que con las series temporales de otras clases.
- Detección de Motifs: Encuentra patrones repetidos en series temporales que no sean previamente conocidas en el banco de datos.
- Pronóstico: Pronosticar eventos futuros con base en eventos pasados conocidos.

D. Descubrimiento de motifs

La búsqueda de motif es una tarea bien conocida en el área de bioinformática. Ese problema también despertó el interés de comunidades de minería de datos [16].

Los motifs son los patrones previamente desconocidos que ocurren con mayor frecuencia en los datos. Esos patrones pueden ser de particular importancia para otras tareas de minería de series temporales, tales como, identificación de agrupamientos, descubrimiento de reglas de asociación, identificación de anomalías y análisis del comportamiento. Un algoritmo eficiente para el descubrimiento de motifs también puede ser útil como una herramienta para sumarización y visualización de grandes volúmenes de datos [8].

Las definiciones siguientes presentan un resumen de la terminología, definida inicialmente en [16], usada para la definición del problema.

- Banco de series temporales: Una base de datos de series temporales S es un conjunto no ordenado de N series temporales posiblemente de diferentes longitudes.
- R-Motif: Para definir el R-motif de un banco de datos de series temporales S, un parámetro de tolerancia R es usado para decidir si dos series temporales son suficientemente similares para ser consideras parte del motif.
- k-Motif: Para definir el k-motif de un banco de datos de series temporales S son usados los k elementos más próximos entre sí.
- Pair-Motif: Esta es una versión del problema, en el que se considera solo las dos sub-secuencias que están más próximas uno del otro, o sea, los que presentan la menor distancia entre sí. Así, el Pair-Motif de un banco de datos de series temporales S es el par no ordenado L1, L2 en S que son lo más similares entre todos los pares posibles.

- Top-KthMotif: El Top-Kth Motif de un banco de datos de series temporales S es el cluster clasificado en la K-esima posición. Así se definen los Top K-Motifs como los K patrones más frecuentes de la base, o sea los primeros K motifs.

A pesar de que muchos trabajos consideran el problema del descubrimiento de motifs como la búsqueda de los motifs en subsecuencias de una serie temporal, esta definición tiene algunas cuestiones a resolver. Por ejemplo, se necesita definir si dos subsecuencias seguidas son lo suficientemente diferentes como para ser consideradas dos subsecuencias independientes, tal como puede ser visto en [16]. Así por cuestiones de simplicidad se considera trabajar solo sobre bancos de datos de series temporales [8].

Intuitivamente, la noción de R-Motif, puede ser vista como el cluster más denso en una proyección 2D de las series. La Figura 1 representa esta idea. La Figura 1, (a) ilustra las series temporales a ser analizadas, en (b) representa el motif encontrado y (c) representa la proyección de las series. Tenga en cuenta que si el motif es definido usando un parámetro de tolerancia R puede ser difícil saber cuál es el radio que mejor define el cluster más denso. Debido a que este parámetro es global y depende del dominio de los datos [8].

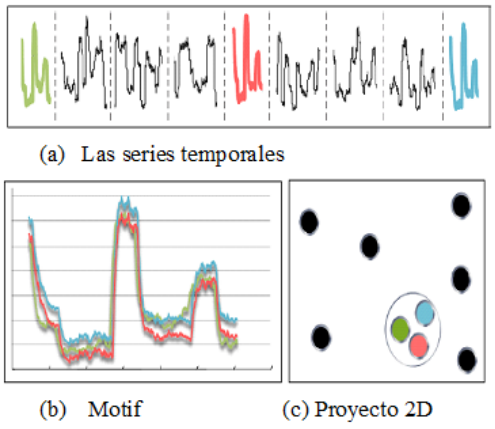

Figura 1. *Representación esquemática del Motif [8].*

Por otro lado, al usar la definición del k-Motif, donde el número de elementos en el motif es definido por el parámetro *k*, este es menos dependiente y más simple de ajustar. Así, el motif es definido por el radio del cluster más denso que contiene los *k* elementos más próximos entre sí [8].

E. Regresión Lineal

Cuando se estudian dos características simultáneamente sobre una muestra, se puede considerar que una de ellas influye sobre la otra de alguna manera. El objetivo principal de la regresión es descubrir el modo en que se relacionan [1].

En la mayoría de los casos la relación entre las variables es mutua, y es difícil saber qué variable influye sobre la otra, pues cada variable influye sobre la otra de forma natural y por igual. El problema de encontrar una relación funcional entre dos variables es muy complejo, ya que existen infinidad de funciones de formas distintas. El caso más sencillo de relación entre dos variables es la relación lineal (2).

$$Y = a + b\,X \text{ (es la ecuación de una recta)} \qquad (2)$$

Donde a y b son números, que es el caso al que nos vamos a limitar. Cualquier ejemplo de distribución bidimensional nos muestra que la relación entre variables no es exacta [1].

F. Naive Bayes

En términos simples, un clasificador de Bayes ingenuo asume que el valor de una característica particular está relacionada con la presencia o ausencia de cualquier otra característica, dada la variable de clase. Por ejemplo, una fruta puede ser considerada como una manzana si es de color rojo, redondo, y aproximadamente 3" de diámetro. Un clasificador de Bayes ingenuo considera cada una de estas características para contribuir de manera independiente a la probabilidad de que esta fruta es una manzana, independientemente de la presencia o ausencia de las otras características [7].

III. Contexto Hidrológico del Área de Estudio

Se sabe que el agua, no solo es un elemento vital para el consumo humano, sino también un factor estratégico para la agricultura, la generación de electricidad, el sector minero e industrial; una gestión eficiente de este recurso impactará directamente en otras áreas estratégicas. El uso de modelos tradicionales de gestión puede ocasionar una deficiente planeación y consecuentemente generar pérdidas físicas, económicas y sociales.

Según el Plan Nacional de Ciencia, Tecnología e Innovación para la Competitividad y el Desarrollo Humano, se tiene el área ambiental priorizada de: Recursos Hídricos y Cambio climático. Además, el apoyo de la tecnología computacional es fundamental y esto se evidencia al ser tratada como área transversal del conocimiento de Industrias de la información y del conocimiento en el desarrollo de software [4].

A. Características de la cuenca Quilca –Chili

La cuenca del río Quilca-Chili se encuentra ubicada al sur del Perú, y su ámbito está comprendido entre las coordenadas geográficas siguientes:

15_37' y 16_47' de Latitud Sur.

70_49' y 72_26' de Longitud Oeste.

Políticamente, se encuentra en la región de Arequipa, abarcando las provincias de Arequipa, Caylloma y Camaná, y algunos pequeños sectores ubicados en las regiones de Puno,

Cusco y Moquegua. El área de la cuenca, hasta su desembocadura en el Océano Pacífico y sin incluir la sub cuenca del Río Siguas, es de 12,542 km2. Sus altitudes varían de los 0 a 6,056 msnm. La cuenca en estudio presenta los siguientes sectores [2] [3]:

Sub cuenca del río Chili (o Sistema Chili Regulado): Este sector, comprende los sitios en los cuales se encuentran las obras mayores de regulación y trasvase (embalses Aguada Blanca, El Fraile, El Pañe y Dique los Españoles y el canal de derivación Pañe-Sumbay), como muestra la Figura 2 y los sitios donde se producen los aprovechamientos del recurso hídrico, como son el uso poblacional, el uso agrícola y pecuario, los usos hidroenergéticos, y los usos mineros e industriales. El área de la ubicación de la Infraestructura Hidráulica mayor del Chili Regulado se encuentra al nor-este de la ciudad de Arequipa, entre las latitudes 15_20' y 16_20' S y las longitudes 71_00' y 71_30' E. Hasta la sección Aguada Blanca, donde se ubica el embalse terminal del sistema, se tiene un área de drenaje de 3,894.9 km2 [4].

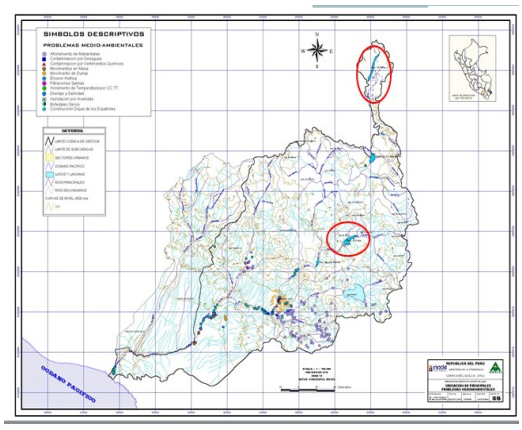

Figura 2. *Embalses Aguada Blanca, El Fraile [18].*

B. Climatología

Evaporación: Para altitudes entre 4,000 m:s:n:m: y 4,600 m:s:n:m, la evaporación anual en tanque, fluctúa entre 1,600 mm y 1,300 mm anuales, respectivamente. En las zonas intermedias, la evaporación anual alcanza 1,825 mm, con una mínima media diaria de 3.1 mm en febrero y una máxima media diaria de 6.2 mm en julio. La evaporación en La Joya, alcanza un promedio anual de 1,752 mm; la mínima media diaria, se registra en abril con 4.3 mm y una máxima media diaria en octubre con 5.6 mm. En las pampas de Majes, estos mismos valores son 2336 mm anuales, 5.5 mm en febrero y 7.7 mm en octubre [2] [3].

Precipitación: Por lo que se refiere a la distribución mensual de la precipitación, se verifica una concentración del 60 al 80%, de la precipitación anual en los meses de diciembre a marzo; en general, el porcentaje es mayor en altitudes menores, lo cual determina también una mayor fluctuación de las descargas durante el año, en cuencas de menor altitud. Los promedios de precipitaciones anuales para estaciones sobre los 4,000 m:s:n:m:, indican valores de 519 mm para Imata (4495 m:s:n:m:), 710 mm para El Pañe (4524 m:s:n:m:), 309 mm para El Fraile (4015 m:s:n:m:). Para altitudes intermedias, se tienen valores de 75 mm para Corpac (2,525 m:s:n:m:), 173 mm para Characato (2,451 m:s:n:m:) y 63 mm para La Pampilla (2,410 m:s:n:m:). Para altitudes como la de las Pampas de La Joya, se tienen valores de 1.8 mm para La Joya (1,255 m:s:n:m:) y para Vítor 17 mm (1,552 m:s:n:m:) [2] [3].

C. Hidrometeorología

Estaciones Hidrométricas: Estación El Pañe:

Se han realizado mediciones, desde 1950 hasta 1964 de las descargas naturales de las lagunas de El Pañe. A partir de 1965, hasta la fecha, en que la presa El Pañe entró en funcionamiento, la estación mide las descargas reguladas, con cortos periodos de interrupción a mediados de la década de los 70. Actualmente, la estación llamada también Oscollo, que es operada por AUTODEMA, está ubicada en el inicio del canal de derivación Pañe-Bamputañe, aproximadamente a unos 100 m de la presa. La sección del canal en este lugar es rectangular, con paredes de concreto de 2.00 m de alto y piso de concreto; su ancho es de 2.70 m y tiene una mira de 2.00 m de alto, ubicada en su margen izquierda [2] [3].

Estación El Frayle:

Se realizaron mediciones desde el año 1953 hasta 1957 de las descargas naturales de El Frayle, luego, dejó de operar, y desde 1964 hasta la fecha, mide las descargas reguladas del reservorio El Frayle, cuya construcción finalizó en 1959 y entró en funcionamiento en 1964. Esta estación de aforos, mide las descargas reguladas por el embalse El Frayle y se encuentra ubicada en el cauce del río Blanco, aproximadamente a unos 50.00 m aguas abajo, del lugar en que ingresan, las filtraciones se ocurren en la represa lateral conocida como Dique de Bloques [2] [3].

D. Exactitud de las mediciones hidrológicas

Al momento de realizar las mediciones dentro de una estación de medición o un punto de control, se debe tener en cuenta que estos valores no son exactos, ya que poseen un error producto del instrumento de medición, el cual no puede eliminarse completamente.

IV. Pruebas y Resultados Obtenidos

Para la realización de las pruebas, se tomó una muestra los registros históricos de estaciones de medición reales: El Pañe y El Frayle, con los datos correspondientes a las medidas de caudal, evaporación y precipitación, tomados entre los años 1970 y 2010.

Muestra: Se tomó los valores promedio por cada mes para el caudal, así como los valores acumulados para evaporación y precipitación.

Para la identificación de Motifs, los datos de entrada son los valores de cada mes desde el año 1970 hasta 2010:

- Caudal
- Precipitación
- Evaporación: Resultado de Caudal y Precipitación

La distribución de datos de Pañe en la Figura 3.

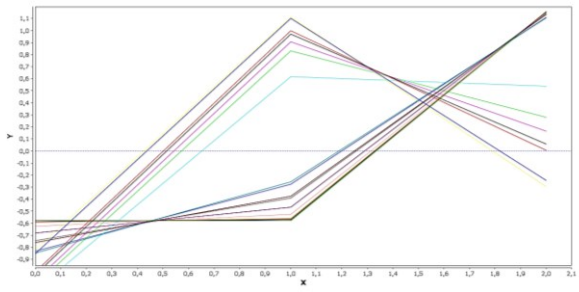

Figura 3. *Distribución de datos de la estación Pañe.*

La distribución de datos de Fraile se observa en Figura 4:

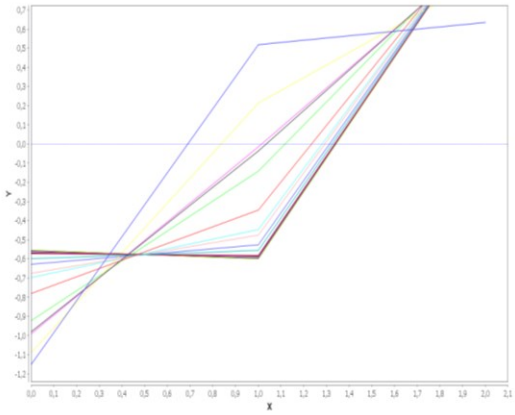

Figura 4. *Distribución de datos de la estación Fraile.*

Mediante la identificación de MOTIFs se encontraron dos cluster los cuales se tomaron como tipos de clasificación en adelante: temporada húmeda y temporada seca

El análisis de la Figura 4, da a conocer dos patrones de comportamiento que se repiten periódicamente, estos patrones se deben a las estaciones del año, por lo que se debe tener en cuenta que existen dos comportamientos muy marcados (Húmedo y seco). A partir de esta información, se considera húmedo si el caudal es mayor a 1 y se considera seco si el caudal es menor a 1.

A. Entrenamiento

Para poder analizar la data, se procede a ingresar todos los datos obtenidos gracias a SENAMHI y a AUTODEMA, teniendo en cuenta que muchos días no fueron recopilados en la base de datos, pero gracias al entrenamiento y a los datos anteriores y posteriores se pudo calcular los datos faltantes mediante el suavizado Medidas Móviles Simples (MMS) [19] que suaviza la observación en el tiempo, calculando una media aritmética de observación y observaciones vecinas [20].

En la Tabla I se muestra un dato al azar por cada mes, que será usado de ejemplo para mostrar la afluencia y básicamente la variación en cuanto a las temporadas de humedad y sequedad gracias a los valores de la precipitación que son los indicadores más relevantes.

TABLA I. EJEMPLO DE ENTRENAMIENTO EN "EL PAÑE" [A]

El Pañe					
Año	***Mes***	***Caudal***	***Precipitación***	***Evaporación***	***HUM/SEC***
1970	1	7.098	183.7	62	HUM
	2	12.431	171.4	92.5	HUM
	3	7.672	130.8	74.4	HUM
	4	2.66	30.1	93	HUM
	5	0.665	12	96.1	SEC
	6	0.359	1.5	96	SEC
	7	0.052	0	96.1	SEC
	8	0.53	1.7	127.1	SEC
	9	0.206	23.3	114	SEC
	10	0.706	26.2	136.4	SEC
	11	0.571	9.1	156	SEC
	12	2.863	126.3	120.9	HUM

B. Pruebas

Los datos se encuentran divididos por las dos clasificaciones encontradas por la clusterización: Humedo y Seco. Para cada división se requiere hallar la fórmula de la recta, por tanto se realizaron 3x2 regresiones lineales:

1. Húmedo (Caudal, Precipitación, Evaporación)

2. Seco (Caudal, Precipitación, Evaporación)

Para realizar la predicción de los componentes del balance hídrico, se analiza cada regresión lineal y se toma una muestra del promedio de años anteriores y las últimas variaciones para dar una estimación aproximada para los 6 meses planteados y estudiar la varianza que podría mostrar los resultados y ayudar en las posibles consecuencias.

1) Pañe:

Una vez obtenidos los datos para cada componente del balance hídrico, y obtenidas las divisiones por la clusterización se realiza la regresión lineal por cada componente tal y como se

observa en la Figura 5 de clúster húmedo y en la Figura 6 de clúster seco.

Como se muestra en la Figura 5 los datos obtenidos de la regresión lineal demuestran la fluctuación en la precipitación en temporada de lluvias, demostrando en comparación a años anteriores, una temporada con lluvias consecutivas que fluctúan dentro de rangos normales, aunque algunos son bajos, indicando que disminuye relativamente los litros de agua.

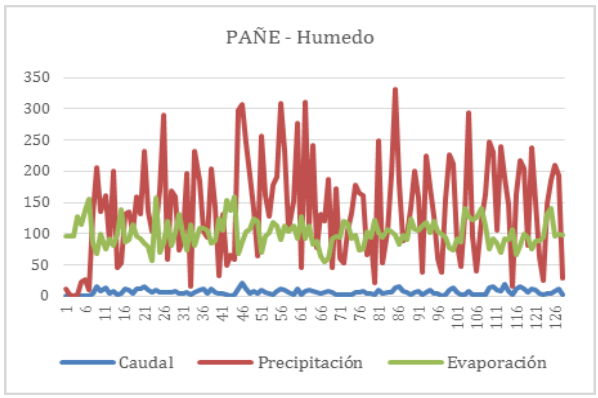

Figura 5. *Regresión lineal represa del Pañe clúster húmedo.*

Seguidamente en la Figura 6 se muestra la conducta que tomarían los componentes del balance hídrico en el resto del año, es decir los valores de evaporación son mas altos que los de precipitación, salvo algunos días donde se observa que habrá un poco de afluencia que se puede aprovechar.

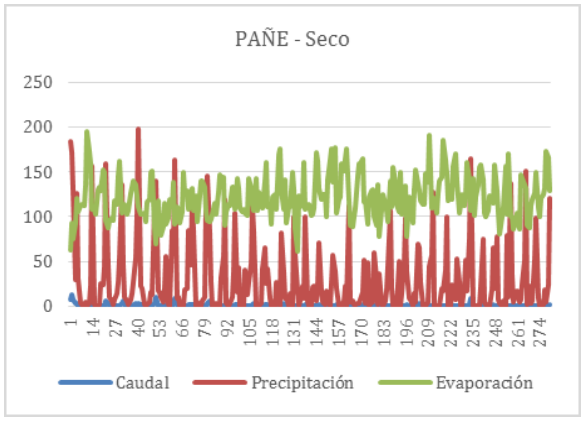

Figura 6. *Regresión lineal represa del Pañe clúster seco.*

2) Fraile:

De igual manera que con la represa del Pañe, se analizó y comparó los resultados obtenidos para la represa del Fraile, como se muestra en la Figura 7, donde la precipitación baja considerablemente en comparación a la represa del Pañe, y además presenta un mayor índice de evaporación.

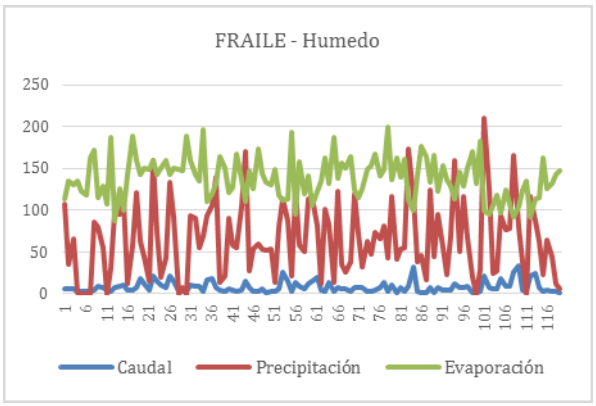

Figura 7. *Regresión lineal represa del Fraile clúster húmedo.*

En el caso del clúster seco la variación es muy notoria, ya que las precipitaciones bajan en un mayor porcentaje y elevándose un poco la evaporación, estos resultados son en tanto menos alentadores ya que indican muy poca afluencia, disminuyendo la probabilidad de posibles precipitaciones por el alto porcentaje de evaporación, de decir el agua acumulada se consume con mayor rapidez, aunque estas pruebas no han sido satisfactorias, pueden aportan a tomar alguna precaución.

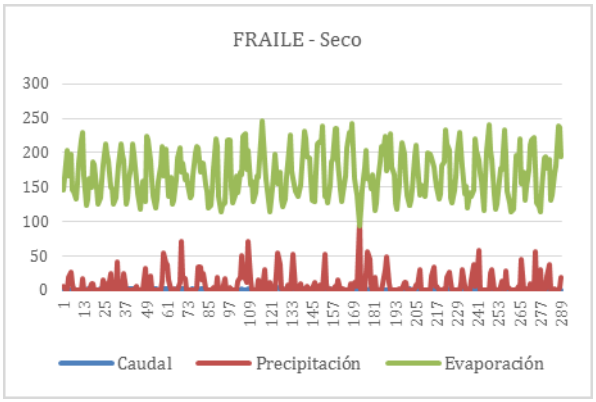

Figura 8. *Regresión lineal represa del Fraile clúster seco.*

Como resumen de los resultados obtenidos del encuentro de motif y clasificación de los datos se muestran a continuación la tabla con los porcentajes de los resultados con regresión lineal por cada variable independientemente, mostrándose los resultados en valor de porcentaje en la Tabla II de acuerdo a lo obtenido para la estación hidrométrica del Pañe como del Fraile.

TABLA II. CUADRO RESUMEN DEL PAÑE Y DEL FRAILE [A].

Resultados de Regresión Lineal		
PAÑE	**Seco**	**Humedad**
caudal	79%	89%
precipitación	77%	96%
evaporación	86%	41%
FRAILE	**Seco**	**Humedad**
caudal	62%	87%
precipitación	23%	80%
evaporación	91%	84%

[A] Resumen de resultados con regresión lineal.

Y por último se tiene una última Tabla III que muestra los resultados con los clasificadores de Redes neuronales multiclase y con naive bayes, para el cual se obtuvo los datos normalizados aleatorios de los resultados dando como los siguientes porcentajes:

TABLA III. RESULTADOS DE CLASIFICACIÓN POR NAIVE BAYES Y REDES NEURONALES.

	Redes Neuronales	**Naive Bayes**
Pañe	97%	95%
Fraile	96%	96%
Predictivo	95%	94%

A través de los resultados de Redes Neuronales y de Naive Bayes, se comprueba que los resultados se muestran con mayor certitud, tanto para la represa el Pañe como para la represa del Fraile.

Para realizar la predicción, se empleó el descubrimiento de patrones mediante motifs descrito en la sección II, el cual dio como resultado, según las Figuras 3 y 4 la identificación de dos grupos, en el primer grupo se observan valores altos, en la columna caudal, al que se etiquetó como húmedo, mientras que a los valores bajos, se etiquetó como seco.

Para observar los cambios ocurridos, según las muestras, en los parámetros húmedo y seco, se visualiza las mediciones en la Figura 9 donde se tiene una muestra de 35 datos para cada mes del año, es decir la serie 1, representa al mes de Enero, la serie 2 representa el mes de Febrero y así sucesivamente hasta el mes 12, Diciembre. Se observa por ejemplo que el mes 12 siempre se mantuvo húmedo mientras que los meses Julio, Agosto y Septiembre, sufrieron variaciones; Usualmente en estos meses se espera un clima seco con aumentos de humedad leve, pero en las últimas muestras se observan cambios muy pronunciados, donde el clima puede verse con picos muy altos representando humedad o picos muy bajos representando sequedad. Esta tendencia de variación se conservó para el cuadro de predicción.

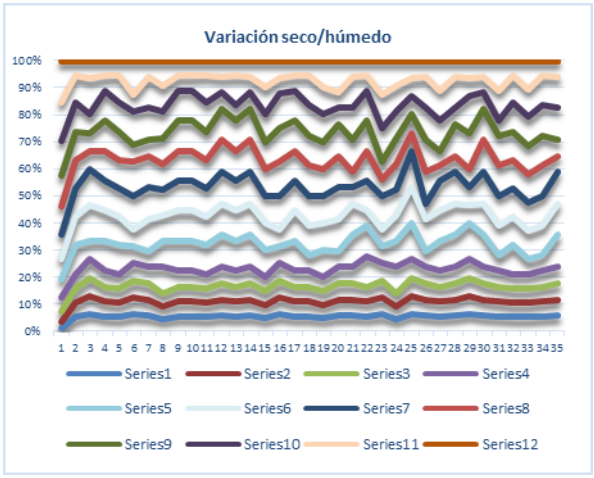

Figura 9. *Variación húmedo/seco de los meses del año para muestra de 35 datos por cada mes de la represa Pañe.*

La siguiente tabla muestra la predicción de 12 meses para la precipitación, evaporación y como resultado de estos datos, obtenemos el caudal.

TABLA IV. DATOS PREDICTIVOS HASTA 12 MESES[B].

Mes	**Caudal Pañe**	**Caudal Fraile**	**HUM/ SEC**
1	6.692088235	5.88853	HUM
2	8.886882353	10.7543	HUM
3	7.244411765	8.70341	HUM
4	2.965323529	2.79247	HUM
5	0.7855	1.20097	SEC
6	0.534676471	1.13612	SEC
7	0.427176471	1.21918	SEC
8	0.602617647	1.13653	SEC
9	0.736852941	1.08126	SEC
10	0.814882353	1.13791	SEC
11	1.095705882	1.10976	SEC
12	2.169882353	1.52985	HUM

[B] Datos agro-meteorológicos de las represas del Pañe del Recurso Hídrico, en la Cuenca del Río Chili, Arequipa.

En la tabla IV se muestran datos predictores, donde la humedad se conserva en los meses: 1, 2, 3, 4 y 12 de la misma forma que se vio en datos de muestra de años anteriores, aunque realizando la comparación con las primeras muestras, se aprecia una tendencia de desplazamiento de este parámetro hacia el mes 5 y va disminuyendo en el mes 12, lo que nos indica los cambios climáticos a través del tiempo que sufre la región. En el resto de meses 6, 7, 8, 9, 10 se sigue conservando el parámetro de

sequedad, y también se tiene la misma tendencia pero, en este caso, hacia el mes 12.

CONCLUSIONES

- Se realizaron predicciones del caudal, gracias a datos tomados desde 1970, donde se halló dos tendencias del clima, el clima seco y húmedo. Donde el clima húmedo es notorio en los meses de diciembre, enero, febrero, marzo y abril; El clima seco predomina en el resto de meses del año.
- Se observó que en los últimos años de la muestra, hubo humedad muy alta o muy baja (seco) en los meses de julio, agosto y septiembre. A diferencia de los primeros años (1970) la humedad en esos meses mantenía un rango menor.
- La utilización de técnicas en minería de datos, como Naive Bayes y Redes Neuronales ayudaron a mejorar la clasificación de los datos y a brindar una predicción con mayor porcentaje de certitud comparando ambos resultados.
- La data meteorológica se trató mediante normalización; Suavizado para filtrar datos erróneos y se completó datos faltantes mediante Medidas Móviles Simples (MMS) para poder obtener una base de datos completa y lograr predecir con mayor porcentaje de certitud.
- Para la sociedad es muy importante la utilización eficiente del recurso hídrico ya que Arequipa es una ciudad en desarrollo, donde la población, la agricultura y la minería, son muy importantes para la toma de decisiones en la distribución y abastecimiento de los mismos. Cabe mencionar que no se contemplaron fenómenos climáticos que hacen variar de forma drástica la tendencia climática.

AGRADECIMIENTOS

Este trabajo ha sido realizado en la maestría de Tecnología de la Información de la Universidad Nacional de San Agustín, la cual es una iniciativa de CITEC a través de un fondo FONDECYT.

REFERENCIAS

[1] Ruíz, D., & Sánchez, A. M. “Apuntes de estadística”. Edición electrónica. Retrieved August, 7, 2013.

[2] Oviedo T., J., Umeres R., H., Franco R., R., Vílchez, G., and Butrón, D. “Diagnóstico de gestión de la oferta de agua de la cuenca quilca – chili”. Technical report, INADE- AUTODEMA, 2001.

[3] OVIEDO, Juan Manuel. “Propuesta de asignaciones de agua en bloque (volúmenes anuales y mensuales) para la formalización de los derechos de los usos de agua en los valles Chili Regulado y Vítor, Chili no regulado e irrigación La Joya del programa de formalización de derechos de usos de agua”, 2004.

[4] EGASA.”Proyecto: Modelado Distribuido de un Sistema Inteligente de Gestión del Recurso Hídrico, Caso: Cuenca del Río Chili, Arequipa”, 2012.

[5] Petrovskiy, M. I. “Outlier detection algorithms in data mining systems”. Programming and Computer Software, 29(4), 228-237, 2003.

[6] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Introduction to data mining Addison-Wesley." 76-79. 4 Independence Way, Princeton, USA, 2005.

[7] Constantino Malagón Luque. Mayo. “Clasificadores Bayesianos, El Algoritmo Naïve Bayes”, 2003.

[8] Mamani, A. V. O. “Soluções aproximadas para algoritmos escaláveis de mineração de dados em domínios de dados complexos usando GPGPU(Doctoral dissertation, Universidade de São Paulo)”, 2010.

[9] Falk, M., Marohn, F., Michel, R., Hofmann, D., Macke, M., Tewes, B., & Dinges, P. “A first course on time series analysis: examples with SAS”, 2012.

[10] Lin, J., Keogh, E., Lonardi, S., e Chiu, B. “A symbolic representation of time series, with implications for streaming algorithms”. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, p´aginas 2–11, New York, NY, USA, 2003.

[11] Agrawal, R., Faloutsos, C., e Swami, A. N. “Efficient similarity search in sequence databases”. In Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms, páginas 69–84, London, UK, 1993.

[12] Keogh, E., Chakrabarti, K., Pazzani, M., e Mehrotra, S. “Locally adaptive dimensionality reduction for indexing large time series databases”. SIGMOD Record, 30(2):151–162, 2001.

[13] K., C. e W., F. A. “Efficient time series matching by wavelets”. In Proceedings of the IEEE International Conference on Data Engineering, p´aginas 126–130, Washington, DC, USA, 1999.

[14] Korn, F., Jagadish, H. V., e Faloutsos, C. “Efficiently supporting ad hoc queries in large datasets of time sequences”. SIGMOD Record, 26(2):289–300, 1997.

[15] Fayyad, U. M., Piatetsky-Shapiro, G., e Smyth, P. “Advances in knowledge discovery and data mining. chapter From data mining to knowledge discovery: an overview”, páginas 1–34. American Association for Artificial Intelligence, 1996.

[16] Lin, J., Keogh, E. J., Lonardi, S., e Patel, P. “Finding Motifs in Time Series”.In 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p´aginas 53–68, Edmonton, Alberta, Canada, 2002.

[17] Muñoz, D. R. “Manual de estadística”. Juan Carlos Martínez Coll, 2004.

[18] Alfaro Casas, L. A. Informe técnico proyecto: “Modelado distribuido de un sistema inteligente de gestión del recurso hídrico, Caso: Cuenca del río Chili, Arequipa”, 2012.

[19] Morettin, P. A., & Toloi, C. M. C. “Time series analysis”, 2004.

[20] Alencar, A. B. “Mineraçao e visualizaçao de coleçoes de séries temporais”. Doctoral dissertation, Instituto de Ciências Matemáticas e de Computação, 2007.