Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Programa de Maestría en Computación

Cubic Spline Interpolation as a distance measure used in the discovery of meaningful rules from complex time series in presence of noise.

"Cubic Spline Interpolation" como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido.

Propuesta de tesis sometida a consideración del Departamento de Computación, para optar por el grado de *Magíster Scientiae en Computación*, con Énfasis en Ciencias de la Computación

Autor:
David Elías Alfaro Barboza

Profesor Asesor:
Luis Alexánder Calvo Valverde

Julio 2016

Cubic Spline Interpolation as a distance measure used in the discovery of meaningful rules from complex time series in presence of noise.

"Cubic Spline Interpolation" como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido.

por

David Elías Alfaro Barboza

Sometida a consideración de la Escuela de Ingeniería en Computación, presentado en Julio 2016, en cumplimiento parcial de los requerimientos establecidos por el *Programa de Maestría en Computación*

Resumen

The ability to make short or long term predictions is at the heart of much of science. In the last decade, the data science community have been highly interested in foretelling about future events by using data mining techniques to find out meaningful rules from different data types, including *Time Series*. Short-term predictions based on *the shape* of meaningful rules may have a vast number of applications in real life. The discovery of *meaningful* rules can be only achieved as a result of algorithms equipped with a robust distance measure, capable to compute precise and noise intolerant similarity results between time series elements. In this work, we believe that *Cubic Spline Interpolation* can be utilized as an efficient distance measure to perform the similarity computation attached into two main algorithms: 1- *"Rule Bit Saves"* and 2- *"Find Antecedent Candidates"*, which were proposed by Mohammad Shokoohi-Yekta et al, to discover meaningful rules from complex time series in presence of noise.

Thesis Supervisor: Luis Alexánder Calvo Valverde
Title: Supervisor

Tabla de Contenido

1 Lista de Tablas 8

2 Lista de Figuras 9

3 **Introducción** 11

4 **Marco Teórico** 12

4.1 Series Temporales 12

4.2 Aplicación y Análisis de Series Temporales 14

4.3 Grandes Retos Sobre la Minería de Series Temporales 15

4.4 Medidas de Distancia para el Cálculo de la Similitud 16

4.4.1 **Distancia Euclideana** 16

4.4.2 DTW (Dynamic Time Warping) 17

4.4.3 **Distancia "Swale"** 19

4.4.4 El Ruido y La Interpolación en Series Temporales . . 19

5 **Propuesta de Proyecto** 20

5.1 Planteamiento del Problema 20

5.2 Propuesta del Proyecto 21

5.3 Trabajos Relacionados 21

5.4 Hipótesis 21

5.5 Métricas 22

5.6 Justificación del Proyecto 22

6 **Objetivo General** 23

7 **Objetivos Específicos** 23

8 **Alcance y Limitaciones** 24

9 **Entregables** 26

10 **Metodología** 27

10.1 Diseño de Experimentos 27
10.1.1 Declaración del Problema 28
10.1.2 Factores 28
10.1.3 Variables de Respuesta 30
10.1.4 Recolección de Datos 30
10.1.5 Análisis Estadístico 31
10.2 Ambiente de Desarrollo 31
11 **Plan de Trabajo** 32

Referencias **33**

Lista de Tablas

1 Listado de entregables, objetivos relacionados y duración 8
2 Cronograma 8

Lista de Figuras

1 Ejemplos de Series Temporales 9

2 Visualizacion de la distancia Euclideana de dos series temporales. . 9

3 Visualizacion de la minimización de una ruta en DTW. 10

4 Visualización del calculo de DTW. 10

1 Lista de Tablas

Tabla 1: Listado de entregables, objetivos relacionados y duración

Entregable	Objetivos	Duración (Dado en Semanas)
Modificación del software utilizado para el hallazgo de reglas significativas	Incorporar todas medidas de distacia en ambos algoritmos	4
Desarrollo e implementación de un ambiente de pruebas	Implementar un ambiente de pruebas que permita la ejecución de las diferentes versiones de ambos algoritmos	3
Documento final con la recopilación y el análisis del diseño de experimentos	Ejecución del diseño de experimentos preparado, incluyendo la medición de las métricas definidas, para determinar si hay diferencias significativas entre los algoritmos.	2
Documento final del análisis de varianza no paramétrico y caracterización de los resultados obtenidos	Ejecutar y reportar el análisis de varianza no paramétrico, para aceptar o rechazar la hipótesis.	2
Documento final de tesis		3
Preparacion de la defensa		2

Tabla 2: Cronograma

Entregable	Semanas															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Modificación de ambos algoritmos para cada una de las distancias																
Desarrollo del ambiente de pruebas																
Ejecución del experimento para cada medida de distancia																
Compilación de los resultados obtenidos en el diseño de experimentos																
Ejecución del análisis de varianza no parametrico y caracterización de los resultados obtenidos																
Documento de tesis																
Preparación de la defensa																

2 Lista de Figuras

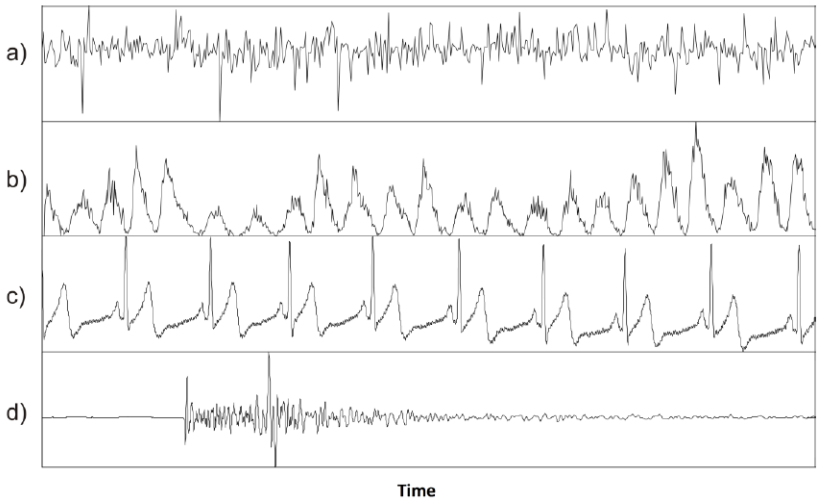

Figura 1: Ejemplos de Series Temporales

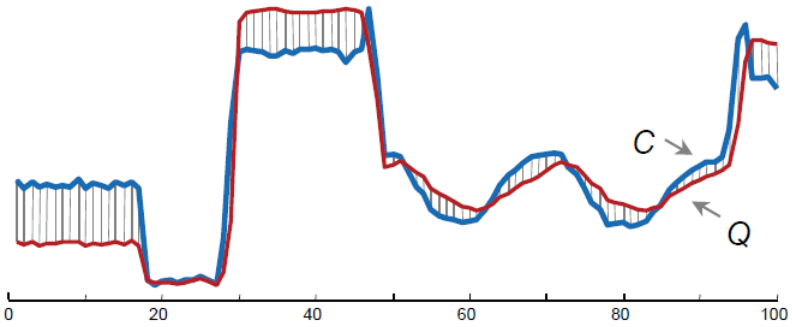

Figura 2: Visualizacion de la distancia Euclideana de dos series temporales.

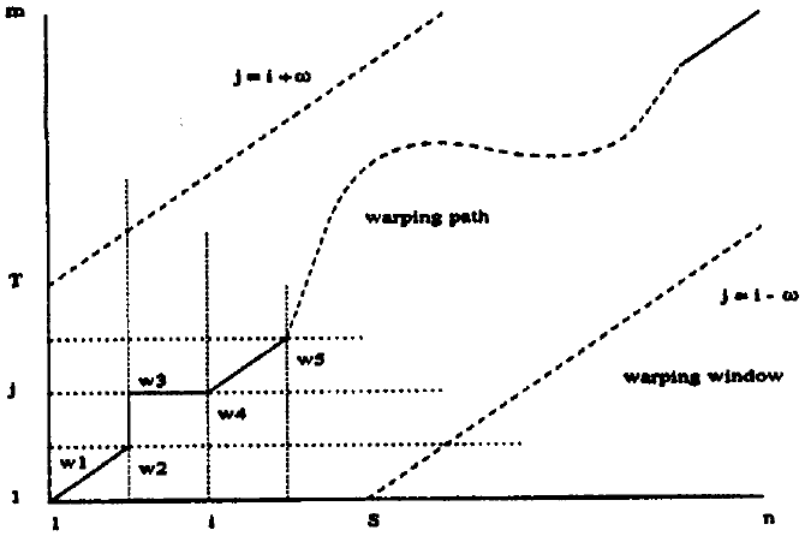

Figura 3: Visualizacion de la minimización de una ruta en DTW.

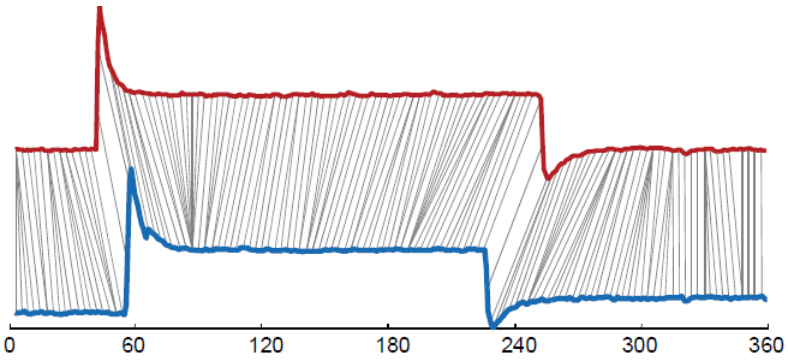

Figura 4: Visualización del calculo de DTW.

3 Introducción

La habilidad de hacer predicciones acerca de acontecimientos o eventos de la vida real ha sido siempre un tema de gran interés para la ciencia.

En la última década, la comunidad de minería de datos se ha interesado vehementemente en el hallazgo de patrones o reglas que puedan ser útiles en la predicción de eventos de corto y de largo plazo [1].

La mayoría de trabajos de investigación recientes, orientados en la predicción de eventos de corto plazo mediante series temporales, se han enfocado principalmente en el análisis de los ***valores actuales*** del flujo de datos [13][18]. Sin embargo, en una basta cantidad de casos, el análisis de los valores actuales es irrelevante; en su lugar, la ***forma*** actual del patrón o la regla motivo y la detección oportuna en el flujo de datos, pueden ayudar a vaticinar el futuro de forma más precisa [1].

Este trabajo de investigación tiene como objetivo principal, la implementación de la medida de distancia llamada *Cubic Spline Interpolation* en los algoritmos ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"***, utilizados respectivamente en el hallazgo y la detección de reglas significativas, para llevar a cabo predicciones de corto plazo sobre series temporales complejas y en presencia de ruido.

Las predicciones de corto plazo sobre series temporales han tenido un auge importante, su aplicación y alcance se ha diversificado considerablemente. Las predicciones de corto plazo sobre texto durante las pulsaciones del teclado, predicciones sobre consultas de base de datos [22], predicciones sobre intervenciones médicas [23], son solo algunos ejemplos de predicciones sobre objectos discretos.

Recientemente, ha surgido una reto aún mayor; se requiere un mayor poder predictivo, lo que implica necesariamente la implementación de algoritmos de predicción mucho más precisos, más veloces y que puedan hallar patrones sobre conjuntos de datos mucho más grandes y complejos [19]. Por ejemplo, el radar Doppler utilizado en las últimas dos décadas para la detección de tornados, ha incrementado el tiempo promedio de alerta de 5.3 a 9.5 minutos, salvando un incontable número de vidas humanas año con año. Sin embargo, aún se reporta alrededor de un 26% de tornados

que no pueden predecirse mediante esta tecnología [20]. McGovern et al. argumentan en [21], que las nuevas mejoras no vendrán necesariamente de sensores más sofisticados, sino, de algoritmos de predicción aún no inventados, que puedan examinar las series de tiempo actuales para hallar reglas predictivas.

Este documento se encuentra distribuído de la siguiente manera: inicialmente, en la sección dos, se desarrolla el marco teórico como un grupo de ideas y conceptos fundamentales que tiene como objetivo principal, guiar al lector en el marco de esta propuesta de investigación. En la segunda sección, se exponen los detalles más importantes de la propuesta de investigación, tales como el planteamiento del problema, la hipótesis, las métricas utilizadas y del por qué la importancia de llevar a cabo esta investigación. El objetivo general y los objetivos específicos de esta propuesta, se ofrecen en las secciones cuatro y cinco respectivamente. El alcance y las limitaciones del proyecto son puntualmente descritas en la sección seis, mientras que los entregables serán enlistados en la sección siete. El diseño experimental y el ambiente de desarrollo comprenden la metodología que se utilizará más adelante en el desarrollo de esta investigación, dentro del tiempo programado en el cronograma de actividades establecido en la sección nueve del documento.

4 Marco Teórico

En la última decada ha surgido una explosión de interés en mineria de datos sobre series temporales. Literalmente, cientos de artículos han introducido nuevos algoritmos para indexar, clasificar, segmentar y agrupar series temporales [24].

4.1 Series Temporales

Las series temporales, pueden definirse como *"una sequencia ordenada de N observaciones (datos) ordenadas y equidistantes cronológicamente, sobre una característica (serie univariable o escalar) o sobre varias características (serie multivariable o vectorial), de una unidad observable, en diferentes momentos"* [25].

Una representación matemática común de una serie temporal *univariable* puede verse de la siguiente manera:

$y_1, y_2, ..., y_N; (y_t)^N t = 1; (y_t : t = 1, ..., N)$, donde y_t es la observación n^0 $t(1 \leq t \leq N)$ de la serie y N es el número de observaciones que componen la serie completa (el tamaño o la longitud de la serie) [27].

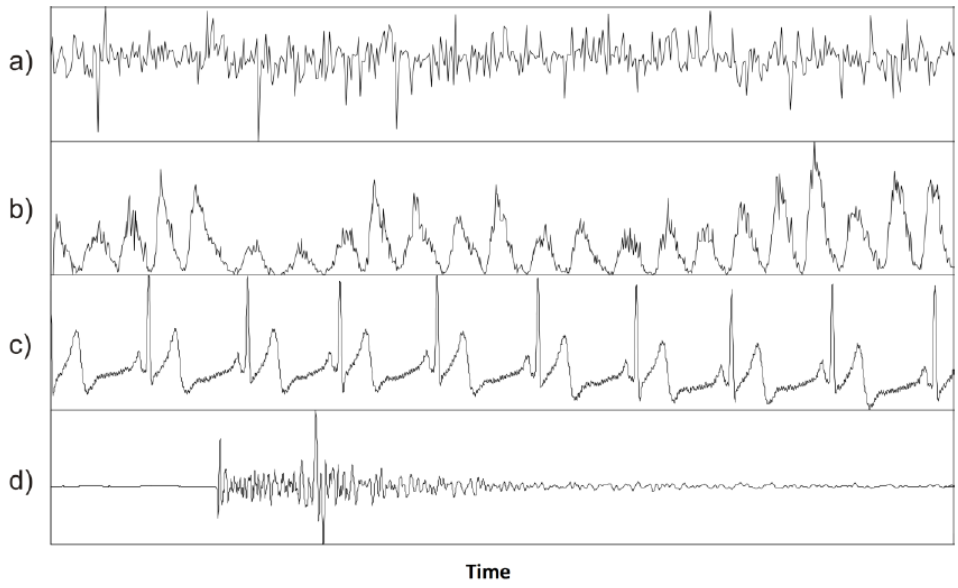

Figura 1. Ejemplo de datos de series temporales relacionados con: **a)** Monzón, **b)** Manchas Solares, **c)** Electrocardiograma, **d)** Señales Sísmicas.
Fuente: [27].

La representación matemática más frecuente de una serie temporal multivariable, puede definirse de la siguiente manera:

$y_1, y_2, ..., y_N; (y_t)^N t = 1; (y_t : t = 1, ..., N)$, donde $y_t \equiv [y_t 1, y_t 2, ..., y_t M]'(m \geq 2)$ es la observación n^0 $t(1 \leq t \leq N)$ de la serie y N es el número de observaciones que conforman la serie completa.

Las observaciones pueden ser almacenas en una matriz Y de orden NXM:

$$Y \equiv \begin{bmatrix} y'_1 \\ y'_2 \\ . \\ . \\ . \\ y'_N \end{bmatrix} \equiv \begin{bmatrix} y'_{11} & y'_{12} & . & . & . & y_{1M} \\ y'_{21} & y'_{22} & . & . & . & y_{2M} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ y'_{N1} & y'_{N2} & . & . & . & y_{NM} \end{bmatrix} \tag{1}$$

donde y_{tj} es la observación n^0 $t(1 \leq t \leq N)$ sobre la característica o variable n^0 $j(1 \leq t \leq N)$, que es la misma en todo momento t [27].

4.2 Aplicación y Análisis de Series Temporales

La medición y el seguimiento del comportamiento de algún fenómeno o los datos específicos de alguna actividad en el tiempo, pueden producir información muy relevante. Es información de primera mano, que puede ser utilizada para comprender mejor aquello que ha venido ocurriendo, monitorear el estado actual para explicar la relación causal o la estructura subyacente que producen los datos observados, y finalmente, permite realizar predicciones o pronósticos que podrían proveer un control anticipado sobre el fenómeno que se está estudiando.

Muchas series temporales tienen una tendencia creciente, por ejemplo, el número de automóviles en uso en casi todos los países durante los últimos cincuenta años, o decreciente como el número de personas que trabajan en la agricultura; otras sin embargo, no tienen tendencia y son estacionarias, por ejemplo, la luminosidad a horas sucesivas, que varía cíclicamente a lo largo de las 24 horas del día.

Por otra parte, existe una amplia gama de aplicaciones, en campos como la medicina, el análisis bursatil, la meteorología, la geofísica, la astrofísica, entre muchas otras disciplinas, cuyas observaciones pueden ser representadas como series de tiempo. Dichas observaciones, pueden ser exploradas y analizadas mediante el uso de técnicas de *mineria de datos*, por ejemplo, a través del descubrimiento de patrones y el agru-

pamiento (clustering), técnicas de clasificación, el descubrimiento de reglas y el resumen o la abstracción de los datos.

Como se explicará más adelante durante el desarrollo de este documento, la propuesta de investigación se enfoca principalmente en la mejora de dos algoritmos utilizados en el descubrimiento de reglas significativas en series temporales.

4.3 Grandes Retos Sobre la Minería de Series Temporales

El análisis y el descubrimiento de patrones sobre series temporales, por definición, presenta una serie de retos y complicaciones que deben abordarse tales como: la super multidimensional, la presencia de grandes cantidades de datos que muchas veces resultan innecesarios o poco útiles durante el análisis, la sensibilidad al ruido o a la presencia de valores atípicos y el dinamismo durante la transmición de datos o *"data streaming"*, debido a que requiere un procesamiento de datos temporales a una gran velocidad, fluctuando constántemente y potencialmente infinitos [1].

Por otra parte, ***el cálculo de la similitud*** durante la comparación de dos o más series temporales es considerado un desafío aún mucho más crucial en la minería de series temporales, especialmente cuando se ha realizado previamente una reducción de la dimensionalidad, de la escala o la amplitud a través del tiempo. La carencia de un alineamiento oportuno sobre el eje tiempo o la amplitud, durante el cálculo de la similitud entre dos o más series temporales, es un problema serio porque tiene un impacto directo sobre el resultado de la comparación. [27].

El cálculo de la similitud, es indispensable en la identificación de segmentos repetitivos contenidos en la serie de tiempo. Estos patrones se conocen como ***"motivos"*** o ***"motifs"***, los cuales, pueden definirse como patrones o *ocurrencias frecuentes de un subconjunto de la serie temporal* [1].

Los motifs pueden considerarse conocimiento puro o significativo, cuando como producto del minado de la serie temporal, se puede utilizar el *motif* para explicar e incluso predecir, con una probabilidad, un fenómeno en el corto o el largo plazo.

4.4 Medidas de Distancia para el Cálculo de la Similitud

Las medidas de similitud son de vital importancia cuando se requieren ejecutar tareas de análisis y minería de datos sobre series temporales, tales como descubrimiento de patrones, agrupamiento, clasificación, descubrimiento de reglas, entre otras. Debido a la naturaleza númerica y continuade de los datos característicos de las series temporales, las medidas de similitud, típicamente se llevan a cabo en forma de aproximaciones.

La complejidad inherente al cálculo de las medidas de similitud, imponen normalmente las principales limitaciones y restricciones de capacidad y tiempo de cómputo, sobre los algoritmos utilizados en el análisis y la minería de datos de series temporales [29]. Es decir, cuanto más rápido y preciso sea el cálculo de la medida de similitud definida en el algoritmo, menor será el tiempo de cómputo necesario para la ejecución del procedimiento completo de minería de datos sobre las series temporales.

Para el desarrollo de este proyecto, se utilizarán específicamente cinco medidas de distancia: 1- Distancia Euclidiana, 2- Swale, 3- Spade, 4- EPR y 5- Cubic Spline Interpolation. Cada una de las medidas de distancia se desarrollará a continuaci'on.

4.4.1 Distancia Euclideana

La distancia Euclidiana es la *distancia "ordinaria" entre dos puntos de un espacio euclídeo* [30]. La distancia Euclideana tiene la característica de ser la medida de distancia más simple y la más utilizada para comparar series temporales [10]. Por ejemplo, si se require comparar dos series temporales Q y C, de largo n, donde $Q = q_1, q_2, ..., q_i, ..., q_n$ y $C = c_1, c_2, ..., c_i, ..., c_n$, se puede utilizar la distancia Euclideana obícua cuya fórmula matemática es la siguiente:

$$DE(Q, C) \equiv \sqrt[2]{\sum_{i=1}^{n} (q_i - c_i)^2} \tag{2}$$

Con se muestra en la **Figura 2**, la visualización del cálculo de la distancia Euclideana puede verse entonces como la raíz cuadrada de la suma de diferencia al cuadrado; tal

y como se observa en cada línea vertical para cada punto de datos desde C hasta Q y viceversa.

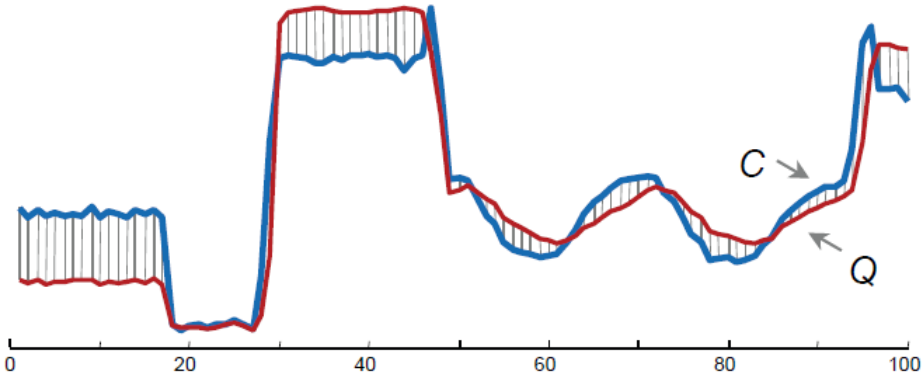

Figura 2. Visualización de la distancia Euclidiana entre dos series temporales C y Q.
Fuente: [30].

4.4.2 DTW (Dynamic Time Warping)

En el año 1994, Berndt y Clifford [32] proponen ***"Dynamic Time Warping"***. En el análisis de series de tiempo, DTW es un algoritmo para medir elásticamente la similitud entre dos secuencias temporales que pueden variar en velocidad y por ende en alineamiento, en relación con el eje tiempo [27].

A diferencia de la distancia Euclidiana, el cálculo de la similitud no se hace en forma lineal, por el contrario, las secuencias son "deformadas" de manera no lineal, para determinar la medida de similitud independiente con respecto a ciertas variaciones en el eje tiempo. Por ejemplo, la tarea de detección de patrones implica la búsqueda de una serie de tiempo S, para instancias de una plantilla T, donde $S = s_1, s_2, ..., s_i, ..., s_n$ y $T = t_1, t_2, ..., t_i, ..., t_n$.

Las secuencias S y T, pueden ser acomodadas para conformar un plano de m por n o un cuadrante, en donde cada punto del cuadrante, (i, j), corresponde a un alineamiento entre elementos s_i y t_i.

Un camino W, alinea los elementos que pertenecen a S y a T, tal que la distancia entre ellos es minimizada.

$$W = w_1, w_2, ..., w_k \tag{3}$$

Es decir, W es una secuencia o un camino específico de puntos en el cuadrante, en donde cada w_k corresponde a un punto $(i, j)_k$.

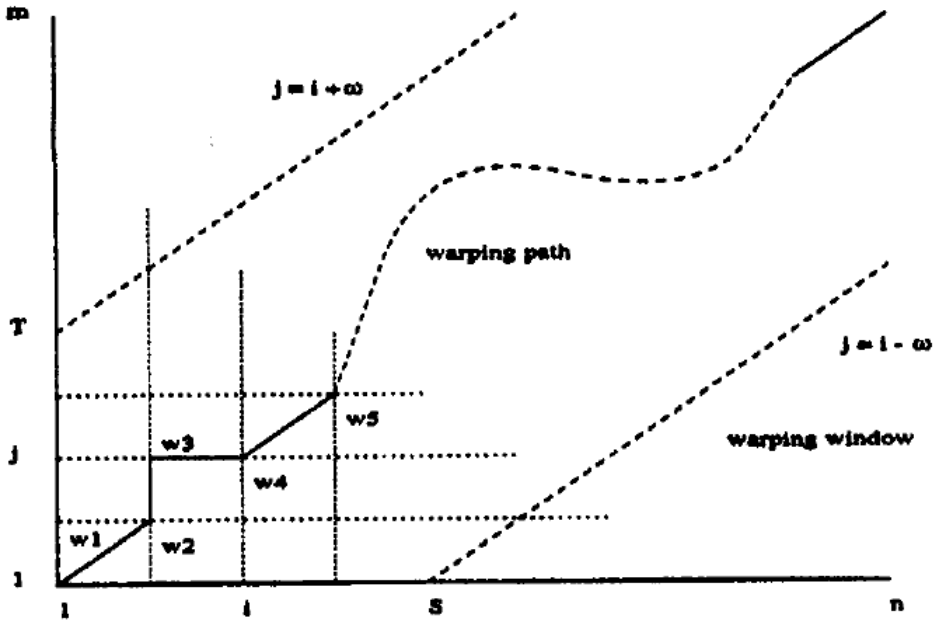

Figura 3. Visualización de un camino w en un cuandrante de m por n.
Fuente: [32].

Posteriormente, por definición, se debe definir una medida de distancia. En [32], los autores proponen $\delta(i, j) = (s_i - t_j)^2$. Una vez definida la medida de distancia, se puede definir formalmente DTW como la minimización sobre los caminos o rutas potenciales no lineales, basados en la distancia acumulada para cada ruta, en donde δ es una medida de distancia entre dos puntos de datos o elementos de las series temporales.

$$DTW(S, T) = min_W[\sum_{k=1}^{p} \delta(w_k)] \tag{4}$$

En la Figura 4, se muestra observa un ejemplo claro de la forma que adoptan las deformaciones no lineales calculadas mediante el algoritmo DTW, para adaptar la diferencia de los puntos de datos con respecto al eje tiempo.

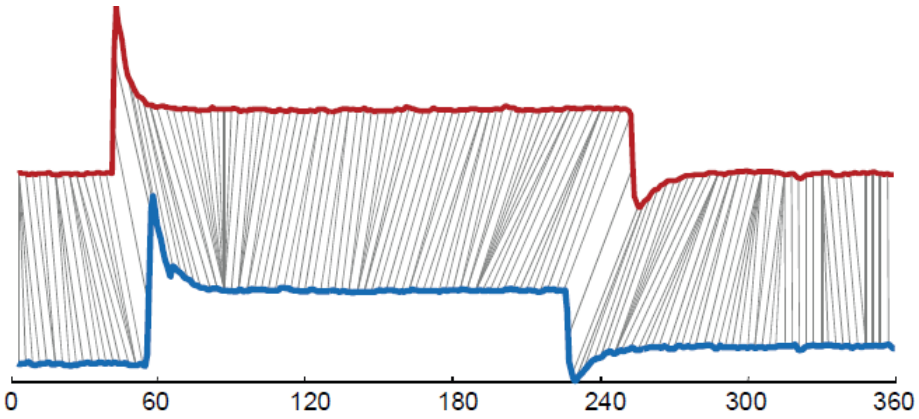

Figura 4. Dos segmentos de series temporales acerca del comportamiento de insectos se alinean con invarianza a la deformación (*"warping"*) ha sido computado mediante el uso de DTW.
Fuente: [32].

4.4.3 Distancia "Swale"

En [31], los autores proponen un modelo de similitud llamado ***"Swale"*** ("*Sequence Weighted ALignmEnt*", por sus siglas en Inglés); este modelo utiliza un sistema de puntuación para recompensar las similitudes y penalizar las disimilitudes durante la comparación de series temporales. El uso de *Swale* requiere necesariamente el ajuste de tres parámetros: 1- un umbral de similitud ϵ, 2- un valor o peso r utilizado para recompensar las similitudes y 3- un valor o peso p para penalizar los vacios o disimilitudes [31].

4.4.4 El Ruido y La Interpolación en Series Temporales

5 Propuesta de Proyecto

5.1 Planteamiento del Problema

El cálculo de la similitud en series temporales ha sido un tema muy estudiado en la última década [13]. La precisión, la velocidad de cómputo y la tolerancia al ruido, son factores claves (especialmente en conjuntos de datos grandes y complejos) a la hora de elegir una medida de distancia robusta para comparar dos series de tiempo de largo n [4].

En la literatura, la medida de distancia más utilizada para comparar series temporarles es sin duda la distancia *Euclidiana* (o alguna de sus variaciones). Dicha medida es principalmente aplicada en el descubrimiento y la comparación de motivos o patrones en series temporales [10][11].

Existen pruebas empiricas fiables, que demuestran que la distancia Euclidiana es muy competitiva e incluso superior a medidas mucho más complejas en una amplia variedad de dominios, particularmente cuando el conjunto de datos se vuelve cada vez más grande [9][15].

Sin embargo, algunos aportes más recientes al estado del arte indican que medidas de distancia conocidas como *Dynamic Tme Warping*, en una o varias dimensiones, puede incluso comportarse de forma más robusta que la distancia *Euclidiana* [5]. Este argumento se soporta principalmente en la sensibilidad conocida que presenta la distancia *Euclidiana* ante la presencia de ruido o pequenñas distorciones observadas al comparar dos series temporales desfasadas con respecto al eje de tiempo [6].

Estimar el grado de ruido en un conjunto de datos es una tarea difícil, lo que sí es seguro, es que la presencia de ruido en series temporales es inevitable o prácticamente inherente [12].

5.2 Propuesta del Proyecto

Apoyados en la premisa anterior, el proyecto pretende estudiar el nivel de acierto obtenido como resultado del descubrimiento de *"reglas significativas"* en series de tiempo, utilizando *Cubic Spline Interpolation* como una medida alternativia de distancia aparentemente superior a la distancia *Euclidiana*, principalmente ante la presencia de distorciones en el conjunto de datos.

Una *"regla significativa"* debe entenderse como un motivo o un segmento repetitivo de la serie de tiempo que puede explicar evento o un comportamiento a partir del conjunto de datos analizado. Una vez identificadas y seleccionadas, estas reglas se pueden utilizar para hacer predicciones de corto plazo sobre el flujo de datos [1].

El proyecto se enfoca en remplazar la distancia Euclidiana y probar que la utilización de otras medidas de distancia (particularmente el uso de *Cubic Spline Interpolation*) pueden ser mucho más tolerantes al ruido y aún así garantizar al menos el mismo nivel de acierto en el hallazgo de reglas significativas en series temporales.

5.3 Trabajos Relacionados

5.4 Hipótesis

Con base en la definición del problema y en la propuesta de proyecto, se define la siguiente hipótesis:

El uso de la medida de distancia Cubic Spline Interpolation, mejora el nivel de exactitud en los algoritmos "Rule Bit Saves" y "Find Antecedent Candidates" propuestos por Mohammad Shokoohi-Yekta y colaboradores, en el hallazgo de reglas significativas en series de tiempo complejas y en presencia de ruido.

5.5 Métricas

El análisis comparativo de los niveles de exactitud obtenidos a partir de la ejecución de los algoritmos según la distancia utilizada, requerirá de las siguientes métricas:

- **Exactitud (Q):**

$$\frac{Total\_Aciertos}{Total\_Predicciones} \tag{5}$$

En el caso más general, se utilizará inicialmente la distancia Euclidiana entre la parte consecuente predicha y las $\boldsymbol{F}$ ubicaciones halladas desde donde la rule fue disparada, un valor denotado como ***"Ferror"***, también conocido como la *media cuadrática*.

Sobre el mismo conjunto de prueba, y mediante el uso del mismo segmento consecuente de la regla, se disparará aleatoriamente $\boldsymbol{F}$ veces y se medirá la distancia *Euclidiana* (Cubic Spline Interpolation y otras), entre el segmento consecuente predicha y la ubicaciones aleatorias F.

Ese valor será denotado como ***Rerror*** (el cual, se promediará entre 1000 ejecuciones aleatorias).

En resumen, la medida de calidad reportada puede definirse como:

$$Q = \frac{Ferror}{Rerror} \tag{6}$$

Los valores cercanos a uno, sugieren que las reglas a prueba no se consideran mejores que encontradas en la estimación aleatoria y los valores significativamente menores a uno, indican que la regla en efecto encuentra una estructura verdadera en los datos. En la mayoría de los experimientos se utilizará un retraso máximo (***maxlag***) de 0 .

5.6 Justificación del Proyecto

6 Objetivo General

Estudiar el nivel de exactitud obtenido al utilizar *Cubic Spline Interpolation* como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido.

7 Objetivos Específicos

Los objetivos específicos de este proyecto son los siguientes:

1. Proponer el uso de *Cubic Spline Interpolation* como medida de distancia utilizada en los algoritmos creados por Mohammad Shokoohi-Yekta y colaboradores (***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"***), para el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido.

2. Realizar un análisis comparativo del nivel de exactitud obtenido al utilizar diferentes medidas de distancia en los algoritmos mencionados.

3. Explicar los resultados obtenidos en el objetivo específico anterior, con el propósito de aprobar o rechazar la hipótesis planteada.

8 Alcance y Limitaciones

El alcance de esta propuesta de investigación se enfoca específicamente en estudiar el efecto de utilizar *"Cubic Spline Interpolation"* como medida de distancia en los algoritmos ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"*** propuestos por Mohammad Shokoohi-Yekta y colaboradores, utilizados para el hallazgo de reglas significativas en series de tiempo complejas y en presencia de ruido.
El diseño experimental incluye también la comparación y el análisis de cinco versiones de ambos algoritmos. Cada versión incorpora la implementación de cinco medidas de distancia distintas: 1- Euclideana, 2- Swale, 3- Spade, 4- EPR y 5- Cubic Spline Interpolation.
Como resultado de esta investigación, se entregarán los siguientes productos:

- La implementación de las cinco versiones de ambos algoritmos para cada una de las medidas de distancia.
- Programas auxiliares para la ejecución y el control del entregable anterior.
- Análisis estadístico para constrastar los resultados de los experimentos.
- Un artículo científico que se entregará al comite editorial de alguna revista o conferencia, con miras a su publicación.

Es necesario delimitar esta investigación por motivos de tiempo y extensión. Es por ello que a continuación se detallan las siguientes limitaciones:

- No se tomarán en cuenta otras medidas de distancia.
- No se utilizará ningún otro algoritmo para la identificación de reglas *motif*.
- No se utilizará ningún otro algoritmo para la detección de segmentos antecedentes de una potencial regla significativa.
- Cualquier otro resultado, documento, software o producción que no se encuentren contemplados en los entregables, no será considerado como parte del alcance de este proyecto.

En resumen, debe considerarse el hecho de que el objetivo principal de esta investigación, se enfoca principalmente en la implemetación de las medidas de distancia sobre ambos algoritmos, para su posterior análisis comparativo, aplicado a cada unos de los cinco diferentes conjuntos de datos; utilizando *"Cubic Spline Interpolation"* como el elemento más importante de la hipótesis planteada.

9 Entregables

Los entregables son los siguientes:

- **Modificación del software utilizado para el hallazgo de reglas significativas:** incorporar las medidas de distancia propuestas para ambos algoritmos.
- **Desarrollo e implementación de un ambiente de pruebas:** este ambiente permitirá ejecutar y medir la exactitud de ambos algoritmos, en el hallazgo y la selección de reglas significativas.
- **Documento final con la recopilación y el análisis del diseño de experimentos:** este entregable implica la ejecución del diseño de experimentos y la construcción de una tabla resumen de los resultados obtenidos.
- **Documento final del análisis de varianza no paramétrico y la caracterización de los resultados obtenidos:** corresponde al análisis de varianza no paramétrico, para aceptar o rechazar la hipótesis mediante la caracterización de los resultados.
- **Documento de tesis:** recopila los entregables anteriores, las conclusiones y el resultado de la revisión de la hipótesis. Incluye además, el desarrollo de un artículo científico.

10 Metodología

10.1 Diseño de Experimentos

Para describir el planeamiento pre-experimental para el diseño de experimentos de este trabajo, (con la información disponible hasta el momento), se usan los *lineamientos* desarrollados en el libro de Douglas C. Montgomery [2]. El esquema del procedimiento recomendado en los lineamientos para el desarrollo de esta etapa incluye lo siguiente:

1. **Reconocimiento y definición del problema:** consiste en desarrollar una declaración clara y sencilla del problema. Una clara definición del problema, normalmente contribuye substancialmente a una mejor comprensión del fenómeno que está siendo estudiado y a la solución final de dicho problema.

2. **Selección de factores, niveles y rangos:** consiste en enumerar todos los posibles factores que pueden influenciar el experimento. Incluye tanto los factores de diseño potencial (los que potencialmente se podrían querer modificar en los experimentos) y los factores perturbadores (los que no se quieren estudiar en el contexto del experimento). También se deben seleccionar los rangos sobre los que varían los distintos factores y los niveles específicos sobre los que se aplicarán las iteraciones del experimento.

3. **Selección de la variable de respuesta:** debe proveer información útil sobre el fenómeno que está siendo estudiado.

4 **Selección del diseño de experimental:** se refiere a aspectos claves del experimento tales como el tamaño de la muestra, la selección del orden adecuado para la ejecución de los intentos experimentales y la decisión de bloquear o no algunas de las restriciones de aleatoriedad en la pruebas.

5 **Llevar a cabo el experimiento:** en esta etapa, es de vital importancia monitorear el proceso cuidadosamente para asegurar la correcta ejecución del experimento con respecto a lo planeado.

10.1.1 Declaración del Problema

Estudiar el comportamiento de *Cubic Spline Interpolation* como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido.
Cubic Spline Interpolation y otras medidas medidas de distancia serán incorporadas en dos los algoritmos específicos llamados ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"*** propuestos por Mohammad Shokoohi-Yekta y colaboradores.
Por cada medida de distancia utilizada, se creará una nueva versión de ambos algoritmos.
La precisión en el hallazgo de reglas significativas será medido a través de la ejecución, la comparación y el análisis de cada versión de los algoritmos, mediante la utilización de al menos cinco fuentes de datos temporales de complejidad variada y en presencia de diferentes niveles de ruido.

10.1.2 Factores

En el diseño de experimentos, un factor es aquel componente que tiene cierta influencia en las variables de respuesta [2].
El objetivo de un experimento es determinar esta influencia. A su vez, cada factor cuenta con varios niveles posibles con los cuales experimentar.
Usando la información recolectada en esta etapa de la investigación, así como la experiencia adquirida por el estudiante y expuesta en los capítulos anteriores, se han seleccionado inicialmente los siguientes dos factores para su estudio:

1. **Las métricas o medidas de distancia**

   Se utilizarán dos algoritmos para la ejecución del diseño de experimental: 1- ***"Rule Bit Saves"*** utilizado en la selección de potenciales reglas o patrones significativos (*detección de reglas Motif [1]*) y 2- ***"Find Antecedent Candidates"***, creado para probar la efectividad de las reglas más significativas sobre cada uno de los conjuntos de datos.

   Ambos algoritmos serán moficados y adaptados a cada medida de distancia y

su ejecución sobre cada conjunto de datos será realizada en forma controlada e independiente.

Las medidas de distancia utilizadas, son las siguientes:

- ***Distancia Euclidiana***
- ***Swale***
- ***Spade***
- ***EPR***
- ***Cubic Spline Interpolation***

2. **El conjunto de datos, tamaño y complejidad**

   Con respecto a la cantidad de ruido encontrado en el conjunto de datos, podemos definir al menos tres diferentes tamaño: 1- conjunto de datos con poco ruido (nivel pendiente de determinar), 2- conjunto de datos con ruido moderado (nivel pendiente de determinar), 3-conjunto de datos con mucho ruido (nivel pendiente de determinar).

   En el experimento, se utilizarán los siguientes conjuntos de datos:

   - 1. **"Energy disaggregation dataset":** contiene el amperaje del consumo diaro de una casa promedio durante un año.
   - 2. **"Zebra finch vocalizations":** contiene las grabaciones del canto de un pájaro *"Zebra"* durante sus primeros 100 dás de vida.
   - 3. **"Daily basis activity data set":** este conjunto de datos contiene información telemétrica de actividades cotidianas de una persona.
   - 4. **"NASA telemetry data":** contiene medidas de voltajes erróneas producidas por las válvulas utilizadas en los transbordadores espaciales de la NASA, utilizadas para el estudio y la detección de anomalías .

	Factores	
Algoritmos	**Conjunto de Datos**	**Medida de Distancia**
Rule Bit Saves Find Antecedent Candidates	Energy Disaggregation Zebra Finch Vocalizations Daily Basis Activity NASA Telemetry Data	Distancia Euclidiana Swale Spade EPR Cubic Spline Interpolation

Imagen 1. Tabla de factores por analizar en esta investigación.

10.1.3 Variables de Respuesta

Dado que la hipótesis afirma maximizar el nivel de exactitud en la identificación de reglas significativas y el hallazgo de segmentos antecedentes sobre los diferentes conjuntos de datos, se ha seleccionado ***la exactitud*** como la única variable de respuesta para ambos algoritmos:

1. ***Medición de la exactitud sobre el algoritmo "Rule Bit Saves"***
   Cálculo del total de aciertos en la identificación de reglas *motif* (potenciales reglas significativas), sobre cada conjunto de prueba, mediante la ejecución de cinco las diferentes versiones del algoritmo.

2. ***Medición de la exactitud sobre "Find Antecedent Candidates"***
   Una vez que las reglas significativas han sido identificadas, se requiere calcular para cada una de ellas, el total de aciertos en el hallazgo de segmentos antecedentes (*predicciones*), sobre cada conjunto de datos, para cada una de las medida de distancia implementadas en las cinco versiones del algoritmo.

10.1.4 Recolección de Datos

Las variables de respuesta serán recolectadas de forma automática una vez concluida la ejecución de cada una de las versiones de ambos algoritmos.
La automatización de la recolección de las variables de respuesta será posible mediante la implementación del ambiente de pruebas.

10.1.5 Análisis Estadístico

Una vez concluída la ejecución del experimento, se deben analizar los resultados para comparar ambos algoritmos en sus diferentes versiones. Para esta comparación, se debe analizar si existe una diferencia significativa en los promedios obtenidos de la variable de respuesta para cada uno de los distintos grupos. Un análisis de varianza permite saber si la diferencia en la media de varias poblaciones es significativa debido a la influencia de alguno de los factores.

Para ese propósito, se utilizará el método estadístico, no paramétrico llamado ***"Kruskal-Wallis-Test"***. Mediante el uso de este método, no se tiene que asumir que las poblaciones siguen una distribución normal, únicamente se utilizarán muestras independientes provenientes de distintas fuentes de datos; poblaciones no relacionadas cuyas muestras no se afectan unas con otras [26].

10.2 Ambiente de Desarrollo

Para el desarrollo de la tesis, se implementará una plataforma que cumpla con dos características principales:

1. Que la plataforma se pueda ejecutar sobre los sistemas operativos Windows y Linux: esto implica que el código fuente debe ser escrito en un lenguaje de programación capaz de correr en los dos sistemas operativos con un número mínimo de cambios y que las bibliotecas utilizadas estén disponibles para los dos sistemas operativos.

2. Que las herramientas y bibliotecas a utilizar sean gratuitas al menos para uso académico.

Basado en estas dos características, se ha elegido una lista de posibles soluciones de software a utilizar. Como se indica, esta lista es preliminar, por lo que se preveen posibles cambios durante el desarrollo de la tesis.

- **Sistema Operativo:** Windows 8.1.
- **Lenguaje de programación:** Matrix Laboratory (MATLAB).

11 Plan de Trabajo

El plan de trabajo de esta propuesta de investigación es por definición la secuencia de pasos que se deberán ejecutar para producir los entregables mencionados en la sección 6 de este documento.

El curso de tesis 2 del ITCR está diseñado para ser ejecutado en un período de *16 semanas*, por lo que es necesario planificar estos pasos de forma tal que sea posible realizarlos en el tiempo esperado, con una alta probabilidad de éxito.

Referencias

[1] Shokoohi-Yekta, Chen, Bilson, Bing Hu, Zakaria and Eamonn Keogh. University of California, Riverside. *"Discovery of Meaningful Rules in Time Series".* KDD 2015, Proceedings of the 21th ACM SIGKDD. International Conference on Knowledge Discovery and Data Mining. Pages 1085-1094.

[2] D. C. Montgomeryx. *"Guidelines for designing experiments, design and analysis of experiments."* 5th Edition, 2000, pp. 13-17".

[3] R. L. Mason. *"Statistical Design and Analysis of Experiments With Applications to Engineering and Science."* Second Edition. John Wiley & Sons. 2003.

[4] M. Vlachos, G. Kollios, and D. Gunopulos. *"Discovering similar multidimensional trajectories."* Proc 18th Int. Conf. Data Eng. pp. 673684, 2002.

[5] S. Chu, E. Keogh, and D. Hart, *"Iterative Deepening Dynamic Time Warping for Time Series"*. pp. 195212, 2002.

[6] H. Li, X. Wan, Y. Liang, and S. Gao. *"Dynamic Time Warping Based on Cubic Spline Interpolation for Time Series Data Mining."* 2014 IEEE Int. Conf. Data Min. Work., pp. 1926, 2014.

[7] C. Ratanamahatana and E. Keogh, *"Everything you know about dynamic time warping is wrong."* Third Work. Min. Temporal Seq. Data, pp. 2225, 2004.

[8] G. Al-Naymat, S. Chawla, and J. Taheri. *"SparseDTW: A novel approach to speed up dynamic time warping."* Conf. Res. Pract. Inf. Technol. Ser., vol. 101, no. December 2003, pp. 117127, 2009.

[9] H. Ding, G. Trajcevski, and P. Scheuermann, *"Querying and mining of time series data: experimental comparison of representations and distance measures."* Proc. VLDB Endow., vol. 1, no. 2, pp. 15421552, 2008.

[10] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, *"Exact Discovery of Time Series Motifs"*. Proc. 2009 SIAM Int. Conf. Data Min., pp. 473484, 2009.

[11] A. M. Denton, C. A. Besemann, and D. H. Dorr, *"Pattern-based time-series subsequence clustering using radial distribution functions."* Knowl. Inf. Syst. Vol. 18, no. 1, pp. 127, 2009.

[12] J. Hu, J. B. Gao, and K. D. White, *"Estimating measurement noise in a time series by exploiting nonstationarity."* vol. 22, pp. 807819, 2004.

[13] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth. *"Rule discovery from time series."* Knowl. Discov. Data Min, pp. 1622, 1998.

[14] G. Al-Naymat, S. Chawla, and J. Taheri. *"SparseDTW: A novel approach to speed up dynamic time warping."* Conf. Res. Pract. Inf. Technol. Ser., vol. 101, no. December 2003, pp. 117127, 2009.

[15] E. Keogh, Exact indexing of dynamic time warping, in: Processings of the 28th VLDB Conference, 2005, pp.358-380.

[16] Michael Morse, Jignesh M. Patel, *"An Efficient and Accurate Method for Evaluating Time Series Similarity". University of Michigan*

[17] Park, S and Chu, S.W. Discovering and Matching Elastic Rules from Sequence Databases. Fundam. Inform. 47, 2001.

[18] Wu, H., Salzberg, B., and Zhang, D., Online Event-driven Subsequence Matching over Financial Data Streams, SIGMOD Conference, 2004: 23-34.

[19] Gribovskaya, E., Kheddar, A., and Billard, A. Motion Learning and Adaptive Impedance for Robot Control during Physical Interaction with Humans. ICRA 2011.

[20] Brotzge, J. and Erickson, S., Tornadoes without NWS warning. Weather Forecasting, 25, 159-172. 2010.

[21] McGovern, et al. Identifying Predictive Multi-Dimensional Time Series Motifs: An application to severe weather prediction. Data Mining and Knowledge Discovery. 2010.

[22] Li, G, Ji, S., Li, C, and Feng, J. Efficient type-ahead search on relational data: a TASTIER approach. SIGMOD Conference 2009: 695-706.

[23] Weiss, S, Indurkhya, N, and Apte, C. Predictive Rule Discovery from Electronic Health Records. ACM IHI, 2010.

[24] Eamonn Keogh, Shruti Kasetty, University of California, Riverside. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration.

[25] Tak-chung Fu, A review on time series data mining, Department of Computing, Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong, 2010.

[26] Yvonne Chan, Roy P Walmsley. Learning and Understanding the Kruskal-Wallis One-Way Analysis-of-Variance-by-Ranks Test for Differences Among Three or More Independent Groups. Published on December 1997.

[27] Carmelo Cassisi, Placido Montalto, Marco Aliotta, Andrea Cannata and Alfredo Pulvirenti. Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining. 2012 Cassisi et al.

[28] Han and Kamber (2005). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, CA.

[29] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under 54 dynamic time warping. In Proceedings of the 18th ACM

SIG KDD international conferenceon Knowledge discovery and data mining, pages 262-270. ACM.

[30] Gustavo Batista, Xiaoyue Wang, Eamonn J. Keogh, A Complexity-Invariant Distance Measure for Time Series, University of California, Riverside, University of Sao Paulo - USP.

[31] Michael Morse, Jignesh Patel. An Efficient and Accurate Method for Evaluating Time Series Similarity. University of Michigan. SIGMOD07, June 1114, 2007, Beijing, China.

[32] D. Berndt, J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In AAAI-94 Workshop on Knowledge Discovery in Databases, pages 359370, 1994.