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

Uso de *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

Junio, 2016

Uso de *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 Junio 2016, en cumplimiento parcial de los requerimientos establecidos por el Programa de Maestría en Computación

Resumen

"El abstract se escribíra aquí"

Thesis Supervisor: Luis Alexánder Calvo Valverde
Title: Supervisor

Tabla de Contenido

0.1 Introducción

Escribir la introducción aquí.

0.2 Propuesta de Proyecto

0.2.1 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 si es seguro, es que la presencia de ruido en series temporales es inevitable o prácticamente inherente [12].

0.2.2 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.

0.2.3 Trabajos Relacionados

0.2.4 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.

0.2.5 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):**

$$\frac{Total\_Aciertos}{Total\_Predicciones} \tag{1}$$

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:

$$Q = \frac{Ferror}{Rerror} \tag{2}$$

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 .

0.2.6 Desarrollo del Proyecto

0.3 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.

0.4 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.

0.5 Alcance y Limitaciones

0.6 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 un 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.

0.7 Metodología

0.7.1 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 esta 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 esta 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.

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 version de version 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.

Factores y Niveles

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 pajaro Zebra durante sus primeros 100 dás.
   - 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.

Variables de Respuesta

Dado que la hipótesis afirma maximizar el nivel de exactitud en la selección y el hallazgo de reglas significativas, se han seleccionado las siguientes variables de respuesta:

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

2. ***Precisión:*** mide la calidad de la medida de distancia para cada version del algoritmo ***"Find Antecedent Candidates"***, en la detección del segmento antencedente de la regla sobre el flujo de datos, para cada conjunto de datos de prueba.

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.

Análisis de Varianza

0.7.2 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).

0.8 Cronograma de Actividades

El plan de trabajo para esta investigación es la secuencia de pasos por ejecutar para generar los entregables mencionados en la sección 6 de este documento.
El curso de tesis 2 del TEC está diseñado para ser ejecutado en un periodo 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.

0.9 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																

0.10 Lista de Figuras

Bibliografía

[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*