APLICACIÓN DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO PARA LA AGRUPACIÓN DE TRAZAS EN EL DOMINIO DE MINERÍA DE PROCESOS

APPLICATION OF UNSUPERVISED LEARNING TECHNIQUES FOR CLUSTERING TRACES IN THE PROCESS MINING DOMAIN

Jaciel David Hernández Reséndiz
Universidad Autónoma de Tamaulipas, México
a2183728004@alumnos.uat.edu.mx

Edgar Tello Leal
Universidad Autónoma de Tamaulipas, México
etello@uat.edu.mx

Heidy Marisol Marín Castro
Universidad Autónoma de Tamaulipas, México
hmarin@conacyt.mx

Gerardo Romero Galván
Universidad Autónoma de Tamaulipas, México
gromero@docentes.uat.edu.mx

Recepción: 22/octubre/2019 **Aceptación:** 4/diciembre/2019

Resumen

La minería de procesos tiene como objetivo el descubrir, monitorear y mejorar los modelos de procesos de una organización a través de la extracción del conocimiento a partir de los datos contenidos en los registros de eventos. En algunos casos, dentro de la tarea de descubrimiento de modelos de procesos, el modelo minado puede ser difícil de comprender e interpretar debido a la diversidad de comportamientos identificados. En este artículo se presenta un enfoque basado en técnicas de aprendizaje no supervisado para la agrupación de trazas para generar modelos más simples y compresibles. Los algoritmos implementados para la agrupación son *K-medias*, jerárquico aglomerativo y agrupamiento espacial basado en densidad de aplicaciones con ruido (DBSCAN). En nuestra propuesta se realiza la sintonización o selección de los mejores parámetros para cada algoritmo de aprendizaje no supervisado, usando la métrica *Silhouette* para mejorar el

agrupamiento de trazas, con lo cual se pueden descubrir modelos de procesos simples con una aptitud media aceptable. Para la validación de nuestra propuesta, las pruebas realizadas se centraron en un caso de estudio del sistema de facturación del hospital AMC, obteniendo al algoritmo jerárquico con mejor desempeño y obtenido una aptitud media de 0.7876.

Palabras Claves: Agrupamiento de trazas, minería de procesos, modelos espagueti, registro de eventos.

Abstract

Process mining techniques aim to discover, monitor and improve the processes performed by an organization through the extraction of knowledge from the data contained in the event logs. In some cases, within the task of discovery of business process models discovered can be difficult to understand and interpret because of the large number of behaviors identified. This article presents an approach based on unsupervised learning techniques for clustering trace to generate simpler and more compressible models. The algorithms implemented for clustering are *K-means*, hierarchical agglomerative and density-based spatial clustering of applications with noise (DBSCAN) algorithms. In our proposal, the best parameters for each unsupervised learning algorithm are tuned or selected, using the Silhouette metric to improve the clustering of traces, with which models of simple processes with an acceptable aptitude can be discovered. For the validation of our proposal, the tests performed focused on a case study of the AMC hospital billing system, obtaining the hierarchical algorithm with the best performance and obtained an average aptitude of 0.7876.

***Keywords*:** Clustering trace, event log, process mining, spaghetti models.

1. Introducción

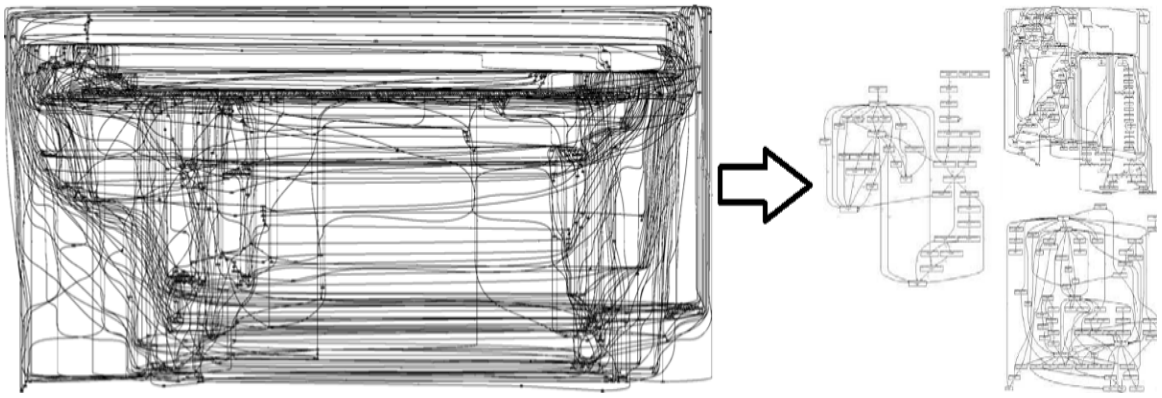
Actualmente, en la mayoría de las organizaciones tienen sistemas de información implementados para gestionar las operaciones de negocio a través de la administración de sus actividades y transacciones. Estos sistemas de información intra o inter-organizacionales permiten realizar tareas en forma automática o semi-

automáticas con la intervención de un usuario. Las tareas, actividades y transacciones pertenecen a un proceso de negocio vinculado a una meta de negocio dentro de una organización [Van, 2016]. Estos sistemas de información, también llamados sistemas de información orientados a procesos [Dumas, 2018], tienen la capacidad de ejecutar procesos de negocio, registrando las acciones realizadas por cada uno de los elementos que componen el proceso de negocio. Estas acciones representan los eventos dentro de un archivo conocido como registro de eventos.

Estos registros de eventos son archivos estructurados con datos sobre las ejecuciones del proceso de negocio. La estructura más simple en estos archivos son los eventos que es la parte atómica de la ejecución de un proceso de negocio específico y que a su vez estos están compuestos por atributos, estos eventos describen un cambio o la ejecución de una actividad. Los atributos contienen información detallada, por ejemplo, el id de la actividad, el usuario que ejecutó la actividad, el dispositivo que ejecutó la actividad, la hora y fecha en que se ejecutó la actividad, entre otros elementos. También, en un registro de eventos se identifican las trazas que es un conjunto de eventos que pertenecen a la misma ejecución de un proceso de negocio. En este sentido, la minería de procesos tiene como objetivo el descubrir, monitorear y mejorar los modelos de procesos a través de la extracción del conocimiento a partir de los datos contenidos en los registros de eventos [Van, 2011], estos objetivos se han abordado en diferentes áreas, por ejemplo, *healthcare* [Rojas, 2016]*, industry* [Van, 2007]*, education* [Van, 2013]*, etc*. Una de las tareas de gran interés dentro de la minería de procesos es el descubrimiento de modelos de procesos de negocio, que consiste en utilizar un registro de eventos como entrada y producir un modelo de procesos de negocio mediante un análisis de los datos contenidos en el registro y la aplicación de un método, tarea y/o técnica de minería de procesos. El descubrimiento permite identificar los comportamientos contenidos en los casos del registro de eventos, con el fin de detectar posibles desviaciones y/o validar que el proceso de negocio se ejecuta de acuerdo con los requerimientos de negocio.

En algunos casos, el modelo del proceso de negocio descubierto por un algoritmo puede ser difícil de comprender e interpretar debido a la gran cantidad de

comportamientos identificados, representado un modelo de proceso de alta complejidad, conocidos como modelos tipo espagueti, tal como se muestra en la figura 1a.



a) Modelo de proceso completo *b) Agrupamiento de trazas*

Figura 1 Modelo complejo de un sistema de facturación del Hospital AMC [Mans, 2008].

Por un lado, se han presentado diferentes trabajos de investigación [Song, 2013], [Koschmider, 2017], [Diamantini, 2014] con propuestas de solución al problema del descubrimiento de modelos de tipo espagueti basados en el agrupamiento de trazas por medio de técnicas de aprendizaje no supervisado. En donde las trazas se agrupan con base a su similitud, con el fin de descubrir diferentes modelos de procesos simples, que en conjunto representen un modelo de proceso complejo. Esta tarea de agrupamiento empieza por la transformación del registro de eventos a un espacio donde es posible calcular la similitud entre trazas. En este tipo de enfoques se utilizan perfiles de características derivadas de la información de los eventos de cada traza. Posteriormente, se aplican técnicas de aprendizaje no supervisado a estos perfiles, generando grupos conformados por las trazas con una similitud cercana.

Dentro de los trabajos reportados en el estado del arte sobre el agrupamiento de trazas por medio de técnicas de aprendizaje no supervisado se encuentra el trabajo [Song, 2013] donde se emplean 3 algoritmos de aprendizaje no supervisado, *K-medias*, Jerárquico aglomerativo y el algoritmo basado en redes neuronales *Self-Organization Map* (SOM). Las pruebas realizadas en este trabajo incluyeron la

ejecución de los algoritmos de aprendizaje no supervisado con 2 perfiles sin procesar y con los perfiles procesados por las técnicas de reducción de dimensionalidad *random projection*, *principal components analysis* (PCA) y *singular value decomposition* (SVD). En el trabajo [Koschmider, 2017] se aplica la agrupación de trazas para descubrir comportamientos similares en el registro de eventos, siendo muy difícil encontrarlos de manera manual en grandes registros de eventos. Por otra parte, en el trabajo [Diamantini, 2014] se aplica un agrupamiento jerárquico para el descubrimiento del comportamiento en un ambiente colaborativo donde se puede tener un conjunto de registros de eventos de diferentes sistemas.

En los trabajos anteriores tienen como limitante la selección o sintonización de los parámetros de los algoritmos de agrupamiento, lo que implica ejecutar en *n* repeticiones los algoritmos con diferentes parámetros para descubrir un modelo de procesos, obteniendo una medida de aptitud media, con el fin de encontrar la mejor agrupación con base a esta métrica. Seleccionando la agrupación con la mejor medida de aptitud media. La aptitud indica el porcentaje en que el modelo de procesos puede representar las trazas del registro de eventos.

Por lo anterior, en este trabajo de investigación se presenta un enfoque basado en técnicas de aprendizaje no supervisado para la agrupación de trazas. Los algoritmos implementados para la agrupación son *K-medias*, jerárquico aglomerativo y DBSCAN basado en densidad. En nuestra propuesta se realiza la sintonización o selección de los mejores parámetros para cada algoritmo de aprendizaje no supervisado, usando la métrica *Silhouette*, la cual permite medir la calidad de las agrupaciones realizadas, facilitando acelerar la selección de los mejores parámetros en el agrupamiento de las trazas con una aptitud media cercana a 0.80, con lo cual se pueden descubrir modelos de procesos simples.

2. Métodos

Agrupamiento de Trazas

A partir de los registros de eventos y la aplicación de técnicas de descubrimiento como los Algoritmos *Alpha*, *Heuristic miner*, *Fuzzy miner* y *Genetic miner* se puede descubrir el modelo de procesos de una organización, algunos de estos registros

de eventos generan modelos de procesos complejos de tipo espagueti. En la figura 1a se muestra un ejemplo de un modelo espagueti derivado del registro de eventos generado por el sistema de facturación del hospital AMC [Mans, 2008], [Van Dongen, 2011].

En el área de aprendizaje no supervisado existen técnicas que permiten agrupar un conjunto de elementos que comparten similitudes. Estos algoritmos pueden permitir agrupar casos o trazas que compartan un comportamiento similar, resultando un conjunto de grupos que estarán conformados por trazas que comparten una similitud y que se pueden usar para descubrir modelos de procesos fáciles de interpretar en comparación con un modelo de procesos tipo espagueti. En la figura 1a se muestra un modelo de procesos descubierto que por su amplitud y comportamiento de los casos es difícil de comprender. Por otro lado, en la figura 1b se muestra un grupo de modelos de procesos que se extrajeron de una parte del registro evento, permitiendo un entendimiento mejor de esa parte del proceso de negocio.

Las técnicas de aprendizaje no supervisados agrupan los elementos con base a una similitud dada por un tipo de distancia (por ejemplo: distancias Euclidiana o Hamming [Pandit, 2011]). Lo cual nos indican qué tan similar es una traza de otras trazas contenidas en el registro de eventos. También, indica que la información de las trazas contenidas en el registro de eventos debe de representarse en un espacio vectorial numérico para determinar la similitud. En este sentido, en los trabajos [Song, 2008] y [Song, 2013] se proponen una serie de transformaciones llamadas “perfiles de trazas” que es una representación de las trazas en un espacio vectorial numérico.

El primer paso para crear estos perfiles es identificar los eventos únicos en el registro de eventos y su origen (tabla 1a, posteriormente se construyen los siguientes perfiles [Song, 2008]:

- Perfil de transición. Para cualquier combinación de dos eventos (A, B), este perfil contiene un elemento que registra la aparición de que un *evento A* ha sido seguido directamente por otro *evento B*. En la tabla 1b se muestra un ejemplo de este perfil considerando la información de la tabla 1a.

- Perfil de actividad. Para cualquier *evento A*, este perfil contiene un elemento que registra la aparición del *evento A* en la traza. En la tabla 1c se muestra un ejemplo de este perfil considerando la información de la tabla 1a.
- Perfil de origen. Para cualquier combinación de un *evento A* y un *usuario X*, este perfil contiene un elemento que mide la frecuencia con la que un *evento A* ha sido realizado por el *usuario X*. En la tabla 1d se muestra un ejemplo de este perfil considerando la información de la tabla 1a.

Tabla 1 Transformación del registro de eventos a los perfiles de trazas [Song, 2013].

a) Información del registro de eventos							
Traza	Información del registro de eventos						
1	(A,Park) (B,Kim) (C,Lee) (D,Choi)						
2	(A,Park) (B,Kim) (E,An)						
3	(A,Park) (D,Kim) (F,Park)						
…	…						
b) Perfil de transiciones.							
	AB	BC	CD	DE	EF	…	
1	1	1	1	1	0	…	
2	1	0	0	0	0	…	
3	0	0	0	0	0	…	
…	…	…	…	…	…	…	
c) Perfil de actividades.							
Traza	A	B	C	D	E	F	…
1	1	1	1	1	1	0	…
2	1	1	0	0	1	0	…
3	1	0	0	1	0	1	…
…	…	…	…	…	…	…	…
d) Perfil de origen.							
Traza	Park	Kim	Lee	Choi	An	…	
1	1	1	1	1	0	…	
2	1	1	0	1	0	…	
3	2	1	0	1	0	…	
…	…	…	…	…	…	…	

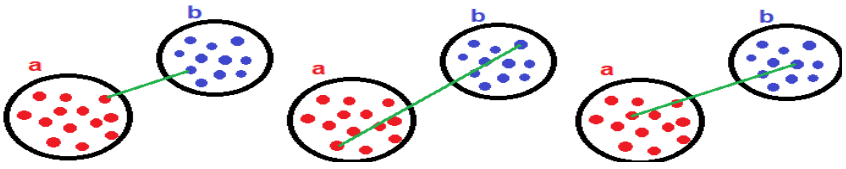
Técnicas de Aprendizaje no Supervisado

En el aprendizaje no supervisado existen tres principales técnicas para la agrupación de elementos:

- Las técnicas de particionamiento se basan en agrupar un conjunto de datos en *K* particiones, donde cada uno de los elementos es agrupado a un único

grupo. Uno de los algoritmos más populares dentro de esta técnica es el algoritmo *K-medias* [Han, 2011].

- En las técnicas basadas en algoritmos jerárquicos aglomerativos [Han, 2011], en donde existen dos variantes:
  - ✓ Ttodos los elementos están agrupados en un único grupo y este se va separando hasta quedar tantos grupos como elementos.
  - ✓ Empieza con tantos grupos como elementos existentes en el conjunto de datos y en cada iteración se agrupan pares de conjuntos o elementos más parecidos. En este algoritmo se basa en el uso de 3 tipos de criterios de enlace para la formación o la separación de elemento (distancia mínima, máxima distancia, y distancia promedio). En la figura 2 se muestra en forma gráfica qué elementos se seleccionan con este tipo de distancias. En este trabajo se usa la segunda variante.
- En las técnicas de agrupamiento basadas en densidad DBSCAN [Aggarwal, 2015], se fundamentan en que es probable que los elementos que están en una zona donde hay una cantidad de elementos significativa o densa deben agruparse en un solo grupo.



a) Distancia mínima *b) Distancia máxima* *c) Distancia promedio*

Figura 2 Distancias para la agrupación de elementos.

Optimización de los Parámetros de las Técnicas Aprendizaje no Supervisado

Una de las principales desventajas de los algoritmos de agrupamiento es definir el número de grupos que se formarán y que están relacionados a los parámetros que reciben. El algoritmo *K-medias* recibe el parámetro *K* que es el número de grupos que se formarán, el algoritmo jerárquico aglomerativo el número de grupos siempre será uno o tantos grupos como elementos existen en el conjunto de datos,

siempre y cuando no se limite el número de grupos formados. Por otro lado, el algoritmo DBSCAN el total de grupos formados dependerá de los parámetros *Eps* y *minPts*, donde *Eps* es el radio considerado para que un elemento pueda ser agrupado con otros elementos y *minPts* es el número de elementos que un grupo debe de tener para que sea válido. Por este motivo, es necesario de apoyarse de métricas que miden la calidad de los grupos formados, con el fin de seleccionar los mejores parámetros que permitan una buena agrupación. Normalmente, estos parámetros son configurados de forma manual. Sin embargo, al desconocer el conjunto de datos, se desconoce el número de grupos y existe el riesgo que el número asignado de forma manual no sea el adecuado para agrupar los datos de manera correcta. Una solución es proporcionar un rango aproximado de valores para los parámetros de los algoritmos y luego usando una métrica para medir la calidad del agrupamiento para determinar los mejores parámetros a partir de la comparación de los resultados de las agrupaciones obtenidas para los diferentes valores de los parámetros.

En nuestra propuesta se emplea el Coeficiente de *Silhouette* [Rousseeuw, 1987] para medir la calidad de los grupos formados y así seleccionar los mejores parámetros para los algoritmos de agrupación. El coeficiente *Silhouette* se refiere a un método de interpretación y validación de consistencia dentro de N grupos de datos. El valor de esta medida identifica qué tan similar es un objeto a su propio grupo (cohesión $a(x)$) en comparación con otros grupos (separación b). La métrica varía de -1 a +1, si el valor es cercano a -1 significa una mala agrupación, si el valor es cercano a 0 la agrupación es indiferente y si el valor es cercano a 1 significa una buena agrupación. Para calcular la métrica *Silhouette* $s(x)$ para un grupo se emplea la ecuación 1 y el coeficiente de *Silhouette* CS para todo el agrupamiento se emplea la ecuación 2.

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{1}$$

$$CS = \frac{1}{N}\sum_{i=1}^{N} s(x) \tag{2}$$

En la figura 3 se muestra la metodología propuesta para la sintonización de los parámetros de los algoritmos de aprendizaje no supervisados en la tarea de agrupamiento de trazas.

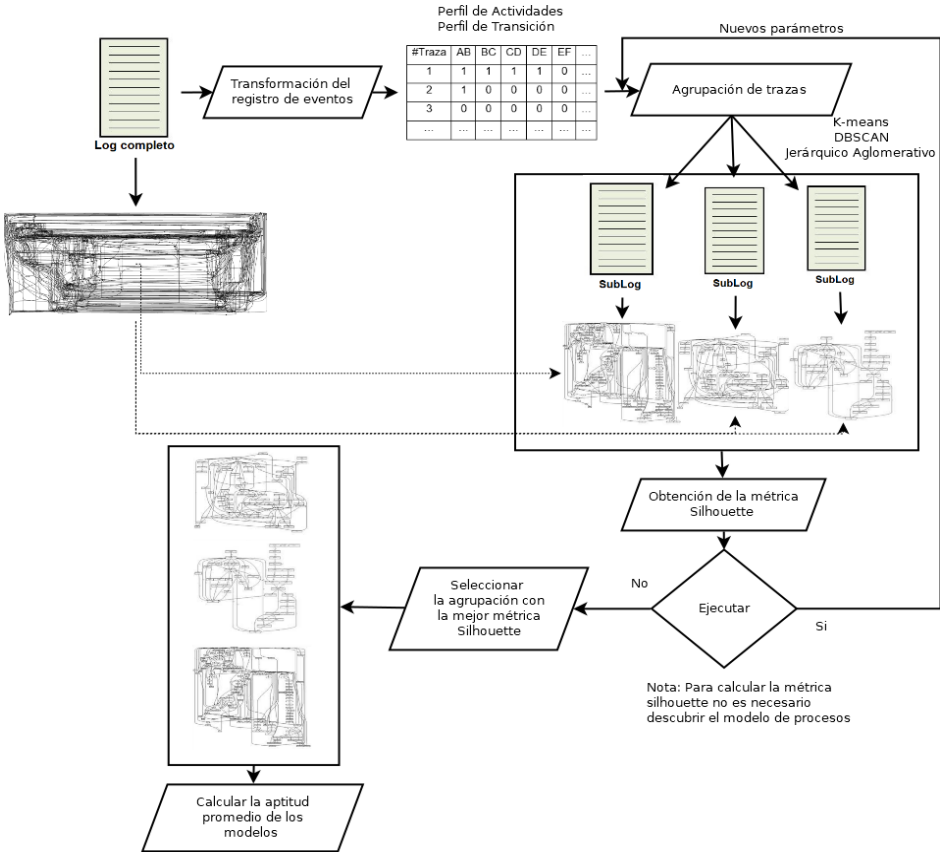


Figura 3 Metodología propuesta para la agrupación de trazas.

Esta metodología se emplea dos veces debido a que utilizamos los perfiles de actividades y de transiciones.

A continuación, se listan las etapas que sigue la metodología:

- El registro de eventos se representa usando un perfil de trazas (de actividades y de transiciones).
- La segunda etapa consta de la ejecución de los algoritmos de agrupamiento con un rango de valores entre 2 y 50 como parámetros, así como los perfiles generados en la tarea anterior. El valor “50” fue seleccionado como valor máximo, este valor es de manera arbitraria porque en realidad se desconoce el mejor número de grupos para el conjunto de datos. Por cada ejecución, los

grupos formados son evaluados usando la métrica de *Silhouette*. Cabe destacar que durante estas ejecuciones no es necesario descubrir los modelos de procesos.

- En la tercera etapa, para cada algoritmo de agrupación se seleccionan los parámetros que obtuvieron la mejor métrica de *Silhouette*.
- A continuación, los parámetros seleccionado en la tercera etapa se usan para ejecutar los algoritmos de aprendizaje no supervisado. Los grupos de trazas formados por estos algoritmos se usan para crear nuevos registros de eventos y descubrir los modelos de procesos más simples. Para el proceso de descubrimiento se usa el *Framework PROM* 6.8 [Verbeek, 2010], que es una herramienta para la minería de procesos ampliamente usada en el dominio de la gestión de procesos de negocio. Mediante esta herramienta se ejecuta el algoritmo *The Heuristic miner* [Weijters, 2006] y el analizador *Replay a Log on Petri Net for Comformance Analysis*, los cuales son aplicados a cada registro de eventos, permitiendo descubrir el modelo de proceso correspondiente y mediante el analizador el modelo se convierte a una red de petri para obtener la aptitud de los modelos de procesos, calculando la aptitud media de las agrupaciones realizadas por los algoritmos.

La media de la aptitud cuantifica el grado en que el modelo de cada agrupación puede reproducir con precisión las trazas agrupadas en la misma agrupación. En otras palabras, el valor de aptitud explica qué tan bien se ajusta un registro de eventos a su modelo de procesos. Si el modelo de procesos puede replicar todas las trazas del registro de eventos, la aptitud sería igual a 1 en otros casos el valor varía entre 0 y 1.

3. Resultados

Caso de Estudio

El registro de eventos generado por el sistema de información del hospital AMC [Mans, 2008], [Van Dongen, 2011] se utiliza en la evaluación de la implementación

de los algoritmos de aprendizaje no supervisado propuestos. El registro proviene de un sistema de facturación del hospital, cada evento se refiere a un servicio prestado a un paciente entre los años 2005 y 2006. El registro de eventos se compone de 624 nombres de eventos diferentes, 1,143 casos y un total de 150,291 eventos.

Escenario de Prueba 1: Selección de los Mejores Parámetros

En el algoritmo *K-medias* se ejecutó variando el valor de *K* entre los valores 2 hasta 50, en la figura 4 se muestra el comportamiento de la métrica *Silhouette* para este rango. En donde *K* = 10 tiene la mejor métrica de *Silhouette* igual a 0.569, usando el perfil de actividades y usando el perfil de transiciones con una métrica de *Silhouette* igual a 0.582.

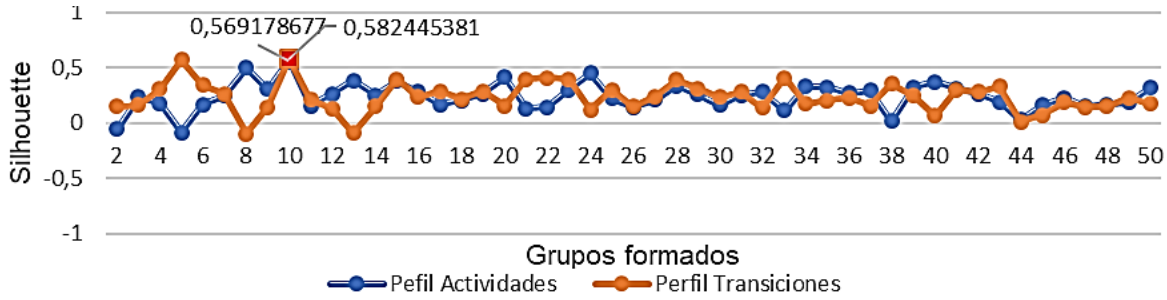


Figura 4 Métrica *Silhouette* para diferentes valores de K en algoritmo *K-medias*.

Para el algoritmo aglomerativo jerárquico se ejecutó variando el total de grupos formados entre 2 y 50, esta misma configuración se ejecutó con los 3 tipos de criterios de enlace distancia mínima, máxima y promedio. La distancia promedio con la mejor medida de *Silhouette* es de 0.73, considerando la formación de 21 grupos y usando el perfil de actividades. La distancia máxima con la mejor medida *Silhouette* es de 0.57. Por otro lado, usando el perfil de transiciones la mejor medida *Silhouette* es de 0.85, usando la distancia promedio considerando 6 agrupaciones. La distancia máxima con la medida *Silhouette* es de 0.84.

Cabe mencionar, que se observa que al usar la distancia mínima en los dos perfiles y al incrementar el número de grupos formados, la medida *Silhouette* mejora. Por este motivo se omite el uso de esta medida, debido a que el número de grupos formados con la mejor medida de *Silhouette* provocará el mismo número de

elementos y la idea que sigue este trabajo es forma grupos de trazas que comparten una similitud. En la figura 5 se muestra el comportamiento de la métrica *Silhouette* para los distintos grupos formados usando el criterio de enlace distancia promedio. Para el algoritmo basado en densidad DBSCAN, los parámetros *Eps* y *minPts* se calcularon usando las distancias de los vecinos más cercanos (*k-nearest*) [Mitra, 2011], [Gaonkar, 2013]. Se calcula el promedio de las distancias de cada traza a sus *K* vecinos más cercanos. El valor de *K* usado es igual a 4 y corresponde al parámetro *minPts*. A continuación, las *K* distancias se trazan en orden ascendente.

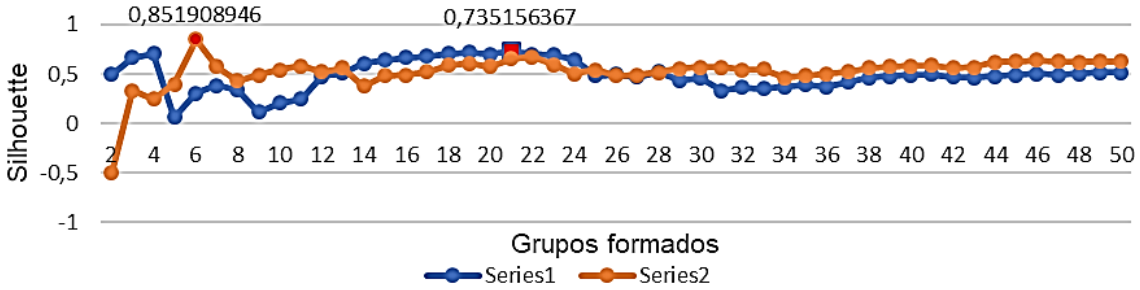
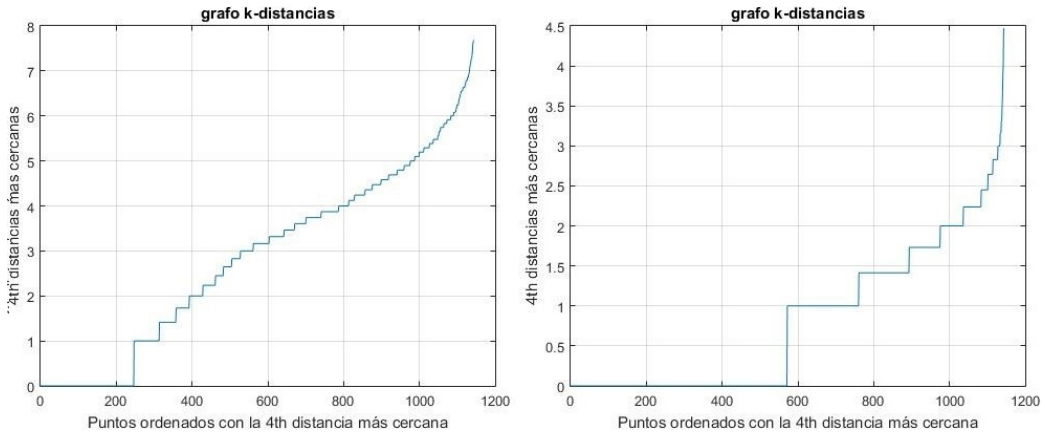


Figura 5 Métrica *Silhouette* para los grupos formados por el algoritmo Jerárquico.

El objetivo es determinar la "rodilla" (cambio drástico), correspondiente al parámetro *Eps* óptimo.

En este sentido, una *rodilla* corresponde a un umbral donde se produce un cambio brusco a lo largo de la curva de *K* distancias. En la figura 6a se muestra la gráfica de las distancias del conjunto de datos para el perfil de actividades. El valor óptimo del parámetro *Eps* se encuentra entre los valores 2.0 a 7.0. Se ejecutó el algoritmo con todos los valores entre estos rangos, sumándole 0.1, produciendo el valor de *Eps* igual a 3.4 con la mejor medida *Silhouette* igual a 0.59, considerando la formación de 5 grupos. Para el perfil de transiciones (figura 6b), el valor óptimo del parámetro *Eps* se encuentra entre los valores 2.0 a 4.5, al igual que el perfil de actividades el algoritmo DBSCAN se ejecutó con todos los valores posibles entre este rango obteniendo el valor de *Eps* igual a 3.7 con la métrica de *Silhouette* igual a 0.52, considerando la formación de 3 grupos. En la figura 7 se muestra el comportamiento de la métrica *Silhouette* para los diferentes valores del parámetro *Eps*.



a) Perfil de actividades *b) Perfil de transiciones*

Figura 6 Selección de los valores adecuados para el algoritmo DBSCAN.

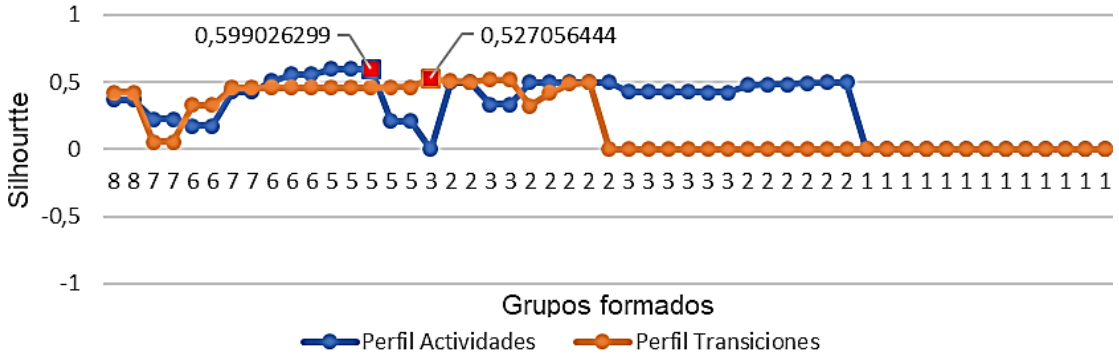


Figura 7 Métrica *Silhouette* para diferentes números de grupos formados por DBSCAN.

Escenario de Prueba 2: Evaluación de los de Procesos Descubiertos

Un medio para validar que los grupos están bien formados se puede realizar mediante el descubrimiento de los modelos de procesos y la obtención de la medida de aptitud media. Por lo cual, se mide la aptitud de cada grupo y se calcula la media de toda la agrupación, esto para todas las agrupaciones realizadas por los algoritmos de agrupamiento considerando los parámetros previamente seleccionados. El resultado de la agrupación de trazas con una combinación que muestre el valor aptitud media más alto se considera la mejor combinación de algoritmo de agrupación y la representación de registro de eventos.

En las tablas 2 y 3 se muestran los resultados de la medida de aptitud para cada uno de los grupos generados por el algoritmo *K-medias* usando el perfil de actividades y transiciones, respectivamente.

Tabla 2 Resultados del algoritmo *K-medias* usando el perfil de actividades.

Grupo	# Trazas	#Eventos	Aptitud		Grupo	# Trazas	#Eventos	Aptitud
1	42	3587	0.6155		6	246	1169	0.6821
2	97	2016	0.4450		7	255	38418	0.7400
3	3	428	0.6780		8	93	699	0.9441
4	197	58055	0.9700		9	119	5996	0.6233
5	89	39678	0.8130		10	2	245	0.8174
						1143	150291	Media: 0.7200

Tabla 3 Resultados del algoritmo *K-medias* usando el perfil de transiciones.

Grupo	# Trazas	#Eventos	Aptitud		Grupo	# Trazas	#Eventos	Aptitud
1	426	3693	0.2848		6	74	12294	0.6490
2	185	40274	0.6374		7	126	25779	0.7902
3	108	33871	0.9300		8	42	5915	0.7806
4	131	16176	0.6775		9	15	7650	0.8307
5	33	4475	0.8007		10	3	164	0.4423
						1143	150291	Media: 0.6823

Los resultados de la aptitud de los grupos formados por el algoritmo jerárquico usando el criterio de enlace de distancia promedio se muestra en las tablas 4 y 5 para los dos perfiles de actividades y transiciones, respectivamente (tabla 1).

Tabla 4 Resultados del algoritmo jerárquico usando el perfil de actividades.

Grupo	# Trazas	#Eventos	Aptitud		Grupo	# Trazas	#Eventos	Aptitud
1	1	990	0.7465		11	1	319	0.7993
2	113	26165	0.6735		12	3	2552	0.8225
3	11	6104	0.7654		13	1	565	0.8000
4	2	1519	0.7266		14	1	902	0.7664
5	913	66833	0.9149		15	1	791	0.7399
6	15	2279	0.6518		16	1	1432	0.8334
7	43	11645	0.7665		17	2	464	0.7385
8	6	4630	0.8344		18	3	2550	0.8039
9	20	15195	0.8901		19	1	964	0.8747
10	3	2152	0.7078		20	1	1814	0.9097
					21	1	426	0.7754
						1143	150291	Media: 0.7876

Tabla 5 Resultados del algoritmo jerárquico usando el perfil de transiciones.

Grupo	# Trazas	#Eventos	Aptitud
1	1	239	0.6437
2	567	116817	0.8888
3	1	553	0.7732
4	569	31525	0.4293
5	4	731	0.7304
6	1	426	0.7754
	1143	150291	Media: 0.7068

De la misma forma, los resultados del algoritmo DBSCAN se muestran en las tablas 6 y 7. La columna “Grupo” indica el número del grupo, la columna “#Trazas” indica el total de casos o trazas dentro del grupo, en la columna “#Eventos” indica es total de eventos dentro del grupo y la columna “Aptitud” se encuentra la aptitud de cada grupo formado. También, al final de cada tabla se encuentra el total de trazas, el total de eventos y la aptitud media del todo el agrupamiento.

Tabla 6 Resultados del algoritmo DBSCAN usando el perfil de actividades.

Grupo	# Trazas	#Eventos	Aptitud
1	678	23008	0.3416
2	7	24	0.3428
3	4	123	0.4323
4	4	140	0.0849
5	1	26	0.3225
	694	23321	Media: 0.3048

Tabla 7 Resultados del algoritmo DBSCAN usando el perfil de transiciones.

Grupo	#Trazas	#Eventos	Aptitud
1	544	121192	0.6920
2	512	12466	0.1689
3	85	15968	0.8090
	1141	149626	Media: 0.5566

4. Discusión

En el enfoque propuesto se analiza el desempeño de los algoritmos de aprendizaje no supervisado para la agrupación de trazas, considerando la selección o sintonización de los parámetros que reciben, con el fin de agilizar el agrupamiento de trazas y el descubrimiento de modelos más simples, con una medida de aptitud media cercana a 0.80.

Cabe destacar que la medida de *Silhouette* permite evaluar la calidad de la agrupación usando la información de los perfiles de cada elemento, ayudando a la selección de los parámetros basados en una métrica para los algoritmos de agrupación. Sin embargo, en una validación desde el punto de vista de minería de procesos se puede utilizar la medida de la aptitud del modelo descubierto de cada grupo formado. En este sentido el algoritmo *K-medias* y el algoritmo jerárquico aglomerativo usando el perfil de actividades se tiene una mejor aptitud media (0.72

y 0.78, respectivamente), tal como se muestra en las tablas 2 y 4. En el algoritmo DBSCAN usando el perfil de transiciones se tiene la mejor aptitud media de 0.55 (tabla 7). Por otro lado, usando el perfil de actividades, el algoritmo jerárquico aglomerativo tiene la mejor aptitud media en comparación con otros algoritmos de agrupamiento, considerando el criterio de enlace distancia promedio y la formación de 21 grupos. El segundo mejor algoritmo es *K-medias* con K=10 y usando el perfil de actividades. Finalmente, el algoritmo DBSCAN presenta un desempeño bajo.

5. Conclusiones

En este trabajo se presentó un estudio de la aplicación de técnicas de aprendizaje no supervisado para la agrupación de trazas en registro de eventos que tienen como característica la generación de un modelo de proceso de tipo espagueti. La metodología propuesta involucra la sintonización o la selección de los parámetros adecuados de los algoritmos de agrupación.

De acuerdo con los resultados obtenidos en la experimentación, se concluye que el uso de la métrica *Silhouette* para seleccionar los parámetros adecuados de los algoritmos de agrupación permite agilizar el agrupamiento de trazas y el descubrimiento de modelos de procesos simples, con una media de aptitud del 0.78, que es el resultado obtenido de la agrupación del algoritmo jerárquico usando el perfil de actividades, considerado con el mejor desempeño de los 3 algoritmos analizados en este trabajo.

Por otra parte, los resultados obtenidos para la aptitud media en la experimentación son cercanos a 0.80, con la posibilidad de alcanzar mejores resultados mediante la aplicación de técnicas de minería de procesos para mejorar los modelos de procesos descubiertos.

En un trabajo futuro, se pretende probar perfiles personalizados que incluyen atributos adicionales de los eventos, por ejemplo: la marca de tiempo, duración del evento, los recursos utilizados, etc. También se pretende agregar en la metodología propuesta, una etapa que permita descubrir las relaciones entre los modelos de procesos simples, con el fin de garantizar el comportamiento del modelo de procesos completo.

6. Bibliografía y Referencias

[1] Aggarwal, C. C. (2015). Data mining: the textbook. Springer.

[2] Diamantini, C., Genga, L., Potena, D., & Storti, E. Discovering behavioural patterns in knowledge-intensive collaborative processes. International Workshop on New Frontiers in Mining Complex Patterns. Springer, Cham, 2014.

[3] Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. Fundamentals of Business Process Management. doi:10.1007/978-3-662-56509-4. 2018.

[4] Gaonkar, M. N., & Sawant, K. AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset. International Journal on Advanced Computer Theory and Engineering, 2(2), 11-16. 2013.

[5] Han, J., Pei, J., & Kamber, M. Data mining: concepts and techniques. Elsevier. 2011.

[6] Koschmider, A. Clustering event traces by behavioral similarity. International Conference on Conceptual Modeling. Springer, Cham, 2017.

[7] Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M., & Bakker, P. J. Application of process mining in healthcare–a case study in a dutch hospital. In International joint conference on biomedical engineering systems and technologies (pp. 425-438). Springer, Berlin, Heidelberg. 2008.

[8] Mitra, S., & Nandy, J. KDDclus: A simple method for multi-density clustering. In Proceedings of International Workshop on Soft Computing Applications and Knowledge Discovery (SCAKD 2011), Moscow, Russia (pp. 72-76). 2011.

[9] Pandit, S., & Gupta, S. A comparative study on distance measuring approaches for clustering. International Journal of Research in Computer Science, 2(1), 29-31. 2011.

[10] Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, pp.53-65. 1987.

[11] Rojas, E., Munoz, J., Sepúlveda, M., & Capurro, D. Process mining in healthcare: A literature review. Journal of biomedical informatics, 61, 224-236. 2016.

[12] Song, M., Günther, C. W., & Van der Aalst, W. M. Trace clustering in process mining. International Conference on Business Process Management. Springer, Berlin, Heidelberg, 2008.

[13] Song, M., Yang, H., Siadat, S. H., & Pechenizkiy, M. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. Expert Systems with Applications, 40(9), 3722-3737. 2013.

[14] Van der Aalst, W., & Mining, W. P. Discovery, Conformance and Enhancement of Business Processes. 2011.

[15] Van der Aalst, W. Process Mining: Data Science in Action, 2nd edn. Springer.New York, USA. 2016.

[16] Van der Aalst, W. M., Reijers, H., Weijters, A., van Dongen, B., De Medeiros, A., Song, M., & Verbeek, H. M. W. Business process mining: An industrial application. Information Systems, 32(5), 713-732. 2007.

[17] Van der Aalst, W., Guo, S., & Gorissen, P. Comparative process mining in education: An approach based on process cubes. In International Symposium on Data-Driven Process Discovery and Analysis (pp. 110-134). Springer, Berlin, Heidelberg. 2013.

[18] Van Dongen, B.F. Real-life event logs - Hospital log. Eindhoven University of Technology. Dataset. https://doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54. 2011.

[19] Verbeek, H. M. W., Buijs, J. C. A. M., Van Dongen, B. F., & van der Aalst, W. M. Prom 6: The process mining toolkit. Proc. of BPM Demonstration Track, 615, 34-39. 2010.

[20] Weijters, A., van der Aalst, W. M. P., & de Medeiros, A. K. A. Process mining with the heuristics miner algorithm. In BETA working paper series WP 166. Eindhoven University of Technology: Eindhoven. 2006.