P-Wave Data Augmentation for Bayès Syndrome Detection

Lorena G. Franco [1[0000-0002-7089-3313]], Luis A. Escobar [2[----------]], Antoni Bayés de Luna [3[0000-0003-1676-207X]] and José M. Massa [4[0000-0002-7456-9676]]

[1] Universidad Nacional del Centro de Bs As, Universidad Nacional Tecnológica, Argentina
[2] Fac. Medicina, Universidad CES, Colombia
[3] Fundación Investigación Cardiovascular. Programa Cardiovascular-ICCC, Institut de Recerca del Hospital de la Santa Creu I Sant Pau, IIB-Sant Pau, Barcelona, España
[4] INTIA, Fac. Cs. Exactas Universidad Nacional del Centro de Buenos Aires, Argentina
francol.edu.ar@gmail.com

Abstract. Resulta interesante detectar en una etapa temprana el Síndrome de Bayés debido a sus asociaciones con múltiples afecciones médicas. En el ámbito de esta investigación se presenta una estrategia de aumentado de datos de muestras de ECGs brindadas por el equipo del Dr Antonio Bayes. Sobre estos datos se aplicaron dos técnicas de clustering: K-Means++ (dos implementaciones diferentes) y FAUM. El método se aplicó mediante la herramienta Matlab y también mediante la provista por FAUM. Además, se utilizó FAUM estableciendo una cantidad fija de clusters. Tanto K-Means++ como FAUM se aplicaron sobre las muestras de cada señal. Inicialmente se contaba con 49 muestras de señales y aplicando las técnicas de aumentado de datos se lograron obtener 2113 señales. Se destaca de los métodos mencionados, la implementación de K-Means++ en el análisis de los agrupamientos. Se logró un F1-Score de 94% en una de sus implementaciones. Los resultados alcanzados son alentadores, ya que el incremento en el conjunto de datos logrado debido al aumentado, hace posible continuar atacando este problema con la aplicación de métodos supervisados que requieran gran cantidad de muestras, como por ejemplo las de aprendizaje profundo.

Keywords: Síndrome de Bayés, ECG, aumentado de datos, agrupamiento.

1 Introducción

En el ámbito de esta investigación se presenta una técnica de aumentado de datos y una comparación de dos métodos de agrupamiento o clustering: K-Means++ (utilizando dos implementaciones diferentes) y FAUM. Los métodos se aplicaron sobre las señales obtenidas con el aumentado de los datos iniciales y fueron utilizados de forma predictiva para clasificar la morfología de la onda P del electrocardiograma (ECG). Cabe aclarar que si bien los métodos de clustering pertenecen a la categoría de métodos no supervisados, como contrapartida de los métodos de clasificación que se categorizan como supervisados, recientemente han surgido aplicaciones de métodos de

clustering con el objetivo de utilizarlos como clasificadores, lo cual se conoce como Classification by clustering [1]. Para esto se han utilizado diferentes enfoques como Weakly Supervision [2], Semi-supervision [3] y otros [4]. En este trabajo se utilizaron las etiquetas de cada clase, no para la construcción del modelo de clustering sino para evaluar la performance de los métodos.

El objetivo se centró en lograr una técnica de aumentado que preserve la calidad de los datos para evaluar su aplicación posterior en la detección del Síndrome de Bayés con técnicas que requieran mayor volumen de datos de entrenamiento. Este Síndrome ha sido estudiado en las últimas décadas por quien le da el nombre, el Dr Antonio Bayés de Luna [5-8] y en los últimos años se ha demostrado su asociación a múltiples afecciones médicas del sistema circulatorio. El concepto de Bloqueo Interauricular (BIA), el más frecuente y relevante a nivel auricular, consiste en la existencia de una conducción retrasada entre la aurícula derecha y la aurícula izquierda. Se dividió el BIA de la misma manera que a nivel ventricular, sinoauricular y auriculoventricular en primer grado o parcial, tercer grado o avanzado y segundo grado o intermitente [6, 8, 9, 10, 11, 12].

Bayés de Luna et al. [5] analizaron un conjunto de ECGs, demostrando una prevalencia de BIA avanzado del 1%, mientras que cuando se seleccionó solo a los pacientes con cardiopatía estructural la prevalencia fue del 2%.

El BIA, que permanece en gran medida subdiagnosticado, presenta notables asociaciones con alteraciones médicas que incluyen fibrilación auricular, isquemia miocárdica, agrandamiento de la aurícula izquierda y émbolos sistémicos [13]. En el artículo [14] se concluyó que el BIA debía ser considerado como un novedoso factor de riesgo para accidente cerebrovascular cardioembólico. En función de lo expuesto resulta de interés su reconocimiento en una etapa temprana.

Desafortunadamente, debido a que el BIA avanzado se presenta en un 1%, a las dificultades de recopilación de datos provenientes de un conjunto de ECGs se le adiciona qué solo puede obtenerse una pequeña cantidad de datos para evaluar el Síndrome de Bayés. En este artículo, se aplica una técnica de aumentado de datos simple para abordar el problema de la escasez de datos en la detección del BIA en el ECG.

El diagnóstico de BIA parcial o avanzado puede realizarse analizando el ECG. En la actualidad no se presentan métodos automáticos de detección del bloqueo mencionado. Sin embargo, cabe destacar la existencia de diferentes métodos de detección de las ondas en el ECG [15-16] y específicamente de la onda P [17-20]. Estos métodos al igual que la mayoría de los recomendados en la bibliografía disponible, abordan con éxito el problema por medio de técnicas basadas en el análisis de frecuencia como Wavelets y Fourier, entre otros. Considerando las múltiples opciones existentes resulta interesante explorar este problema desde el punto de vista de la clasificación. Las técnicas para estos problemas han registrado una importante mejora en su eficacia y eficiencia en los últimos años, impulsadas por los problemas clasificados como Big Data [21]. En el contexto señalado el interés se centró en técnicas de agrupamiento de forma semi-supervisada en las que una muestra de cada clase se etiqueta manualmente.

La onda P revela información valiosa, aunque su detección precisa es una tarea difícil y desafiante debido a su poca amplitud, su baja relación señal ruido, las oscilaciones de la línea de base y morfologías extrañas que pueden presentarse [20, 22].

Hoy en día los avances tecnológicos permiten que muchos equipos generen los resultados del ECG en más de un formato. A pesar de esta evolución, solamente algunos centros y especialistas en Cardiología con gran volúmen de pacientes almacenan los electrocardiogramas en formato digital. Frente a este contexto y debido a que los ECG disponibles son resultado del seguimiento a lo largo de los años de pacientes que presentaron BIA, se procesaron electrocardiogramas que se encuentran en soporte papel y por lo tanto, fue necesaria su digitalización. La misma se realizó teniendo en cuenta que se debe preservar principalmente la onda P. En trabajos anteriores [23] los autores de este trabajo exploraron técnicas de digitalización y segmentación orientadas a preservar esta onda. Respecto a la aplicación de técnicas de clustering, en experimentos previos se exploraron dos de estas técnicas sobre las escasas muestras disponibles, para agrupar ondas P en diferentes grupos. Los resultados obtenidos fueron alentadores [24].

En este trabajo se propone ampliar el alcance de la investigación, trabajando sobre un conjunto de datos obtenidos mediante estrategias de aumentado de datos. Se pretende agrupar ondas P en diferentes grupos correspondientes a morfologías utilizadas en el diagnóstico del BIA: Onda P normal, Onda P bimodal (bloqueo de primer grado), Onda P con morfología negativa y por último onda P bifásica o ± (bloqueo de tercer grado).

Como se puede observar en el estado del arte presentado, los problemas de análisis de señales temporales han sido tratados en general por medio de técnicas de análisis de frecuencias. Sin embargo, en las últimas décadas, el avance del área de conocimiento de la Ciencia de Datos ha contribuido al desarrollo de métodos de agrupamiento y clasificación. Dentro de los primeros, el método K-Means (y sus variantes más modernas como K-Means++) [25-26] ha demostrado ser exitoso en una gran cantidad de problemas [27]. Este método agrupa datos en una cantidad fija k de clases. El otro método utilizado en este trabajo se basa en una técnica novedosa llamada Fast Autonomous Unsupervised Multidimensional (FAUM) [28]. Esta última técnica incorpora un algoritmo heurístico basado en un análisis de la entropía y el equilibrio de cardinalidad entreclusters. FAUM encuentra de forma automática una cantidad de clases de forma no supervisada. Además, es posible establecer una cantidad fija de clases al igual que K-Means, aunque internamente utiliza un algoritmo jerárquico basado en la cardinalidad de clases y el uso de funciones de distancia no euclídeas en casos de dimensiones mayores a 3.

Como se mencionó anteriormente, si bien estos métodos pertenecen a la categoría de métodos de agrupamiento no supervisado, como se observa en la bibliografía [29-30], es posible utilizarlos como clasificadores si se los utiliza de forma semi-supervisada, introduciendo en cada clase una muestra etiquetada de forma de etiquetar el resto de las muestras de la misma clase luego del agrupamiento.

Luego de aplicar estos métodos, se observó que es posible agrupar las muestras conteniendo ondas P en los grupos correspondientes. Esto puede evidenciarse a través

de los indicadores de análisis de la matriz de confusión: Precision, Accuracy, Recall y F1-Score.

A continuación, en la sección 2 se presentará una síntesis de los materiales y métodos de agrupamiento propuestos. En la sección 3 se muestran los resultados obtenidos. En la sección 4 se presentan las conclusiones. Por último se encuentran la bibliografía.

2 Materiales y Métodos

En esta sección se presentan las características de los ECG utilizados y los métodos de agrupamiento aplicados.

2.1 Materiales

En cuanto a los materiales, se partió del conjunto inicial de 49 muestras correspondientes a un total de 600 ECG procedentes de las investigaciones del Dr. Bayés y su grupo de trabajo, utilizadas en trabajos anteriores [23-24]. En particular se aplicó la digitalización presentada en [23] para preservar la onda P. En la Fig. 1 se observan ejemplos de ondas P.

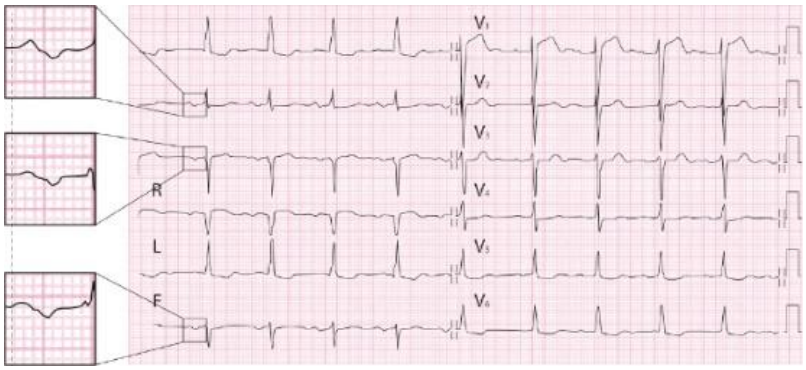
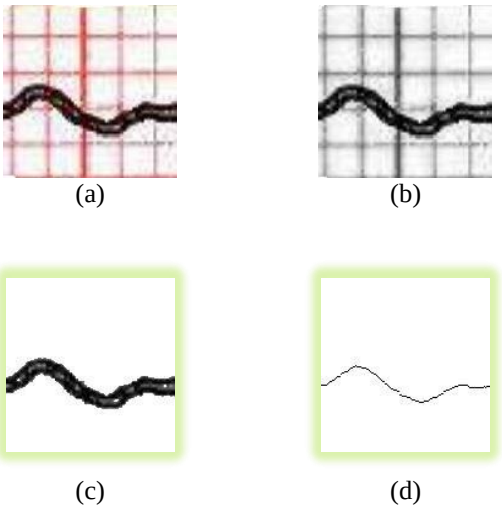


Fig. 1. ECG con BIA Avanzado.

Las imágenes digitalizadas constituyen los materiales de este trabajo. El proceso de digitalización se realizó por medio de un umbralado y una binarización. Con la intención de obtener una curva del menor espesor posible, y por lo tanto mejorar la precisión, se aplicó una esqueletización por medio de una erosión múltiple iterativa. Este proceso se ilustra en las Fig. 2a., 2b. , 2c. y 2d. Con el fin de establecer una referencia para los valores positivos y negativos de la onda, se aplicó un método basado en la técnica manual utilizada por los médicos que trabajan en este tema [31]. Por último se obtuvo una lista de valores de intensidad para cada columna de la imagen, correspondiente a cada elemento de muestreo temporal del ECG. Esta lista de valores se calculó inicialmente con una alta precisión en punto flotante y luego se normalizó entre -1 y 1.



(a) (b)

(c) (d)

Fig. 2. Digitalización de imagen. (a) Imagen original. (b) Binarizado. (c) Umbralización. (d) Esqueletización.

A continuación se aplicó la técnica de aumentado de datos que se presentará en la sección 2.2. Se obtuvieron un total de 2113 señales que a su vez consituyen los materiales para la aplicación de las técnicas de clustering que se presentarán en la misma sección.

2.2 Métodos

En esta sección se presentan los métodos que se aplicaron inicialmente para lograr el aumentado de las señales, luego la construcción del conjunto a agrupar con las señales aumentadas, el agrupamiento por K-Means++ (implementación con Matlab y FAUM) y finalmente el Agrupamiento por FAUM ajustando la cantidad de clusters.

Aumentado de señales

En trabajos previos [23-24], uno de los mayores inconvenientes, fue la limitación en la cantidad de señales disponibles para poder no solo llevar a cabo la clasificación, sino también para tener una evaluación más precisa de la performance de los métodos aplicados. En el marco de esta investigación, se propone cómo solución técnicas de aumentado de datos [32-33].

En la búsqueda de aumentar la diversidad de datos disponibles se consultó a expertos y bibliografía sobre los valores máximos y mínimos de la onda P, tanto para morfologías positivas como negativas. Específicamente respecto de las derivaciones que se estaban analizando: II, III y AVF, actualmente sólo se encuentra definido, para las derivaciones de interés, que la onda P normal positiva puede tener un alto máximo de 2,5 mm [31], [34]. En función de este valor y teniendo en cuenta que las imágenes de las ondas P originales eran de 138x138 pixeles representando 5x5 mm, se determina (considerando el eje), el valor máximo de la onda P en píxeles. La diferencia entre 69

(2,5 mm) y el máximo de la onda P indica la cantidad de imágenes nuevas que se pueden obtener. En cada imagen el valor máximo de P se sube un pixel hasta alcanzar el valor máximo de la onda P normal (2,5 mm). El resto de los valores de la curva positiva se modifican de manera proporcional en base al valor máximo. En relación a la cantidad de datos disponibles en las muestras en papel y al proceso de digitalización, la cantidad de valores de amplitud para cada señal es de 138, lo cual constituirá la cantidad de características para los métodos de clustering que se presentarán en las siguientes secciones. Los valores de la curva que están por debajo del 10% del valor máximo se mantienen sin cambios (ver Fig. 3). Cada imagen nueva obtenida con esta técnica requiere ser ubicada en el centro de la imagen.

La técnica aplicada tiene como objetivo construir datos sintéticos mediante la transformación de muestras existentes. Previamente a la aplicación de los métodos de agrupamiento, las imágenes se normalizan. En la Fig. 4 y Fig. 5 pueden observarse dos señales normalizadas correspondientes a la imagen de la Fig.3, en el caso de la Fig. 4 con valores que van de -1 a 1 para adaptarse al procesamiento en punto flotante que realiza Matlab y en el caso de la Fig. 5 con valores entre 0 y 255 para adaptarse al procesamiento en tipos enteros que realiza FAUM. Cuando se realiza el agrupamiento de la onda P de las imágenes disponibles, es probable que al aumentar sólo la onda positiva de una imagen no cambie su clase. Las muestras sintéticas se agregarán a los conjuntos de imágenes originales para enriquecer las pruebas de clustering.

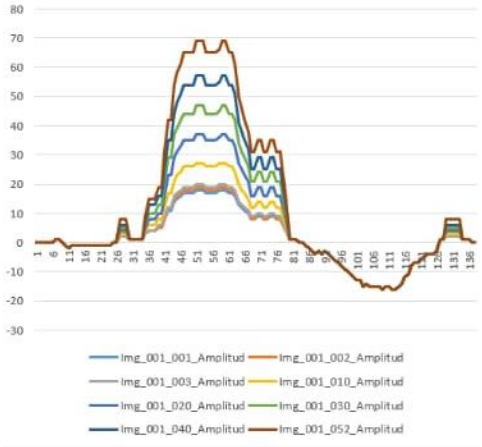


Fig. 3. Señal ampliada

El volumen y la diversidad de los datos son esencialmente importantes al momento de obtener conclusiones de un modelo robusto de agrupamiento. En este artículo se pretende analizar el beneficio de aumentar los datos con muestras creadas sintéticamente cuando se aplican diferentes métodos de agrupamiento.

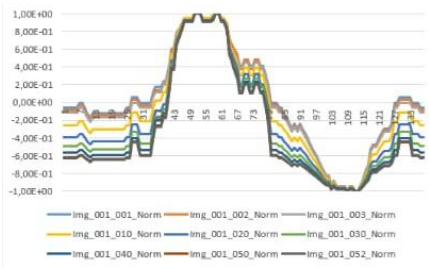


Fig. 4. Señal normalizada para aplicar Matlab

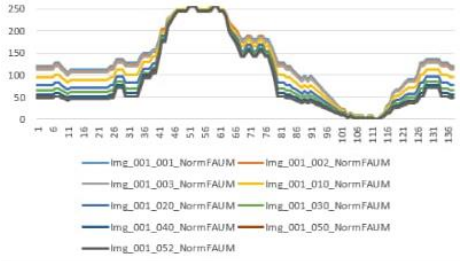


Fig. 5. Señal normalizada para aplicar FAUM

Construcción del conjunto a agrupar

Para poder aplicar los métodos de agrupamiento, en este caso K-Means++ en Matlab, K-Means++ en FAUM y FAUM con cantidad fija de clusters, fue necesario construir un conjunto de datos con las 2113 muestras y para cada una de ellas el conjunto de características a utilizar por los métodos. Es importante mencionar que las 138 características (que representan la amplitud de la señal en cada uno de los períodos de tiempo de medición) provienen del proceso fisio-biológico que consituye el sistema cardiovascular del paciente, el cual es un fenómeno en cierta manera integral, por lo tanto las características poseen una fuerte dependencia temporal entre ellas (en términos simples significa que la amplitud de la onda en un cierto momento de tiempo tiene una fuerte relación con los valores anteriores). Esto significa que la suposición de independencia u ortogonalidad de las características que es deseable en la aplicación de métodos de Machine Learning y en particular en los métodos de clustering [35-37], no se cumple en este caso. Sin embargo, lo que sucede en el problema planteado en este trabajo, es que si bien el comportamiento en este caso de la onda P tiene una cierta predecibilidad en cuanto a lo que se espera que sea una onda P, existen fenómenos fisiológicos, de captura de señal y del mismo Síndrome de Bayés que alteran la forma de la onda. Es por esto que se decidió trabajar con los valores de amplitud temporales como si fuesen independientes para los algoritmos, aunque no lo sean realmente.

El resultado del conjunto, obtenido a partir de las características de la señal, es entonces una matriz de valores, cuyas filas contienen las 2113 muestras y las columnas los 138 valores de amplitud correspondientes a cada lapso de tiempo muestreado.

Estas columnas se corresponden con las características que utilizan los distintos métodos de agrupamiento. Específicamente se construyeron dos conjuntos, el conjunto *Mf*, conteniendo valores de punto flotante entre -1 y 1 para los valores de amplitud y el conjunto *Mi* conteniendo valores enteros entre 0 y 255. Los valores del conjunto *Mi* han sido normalizados debido a que el método FAUM se basa en operaciones de desplazamiento de escalas de magnitud en representaciones de punto fijo.

A continuación, en las próximas secciones de este apartado, se presentan los dos métodos de agrupamiento aplicados al conjunto de datos generados.

Agrupamiento por K-Means++ (Implementación de Matlab y FAUM)

Se aplicó el método de K-Means++ implementado en MATLAB utilizando el conjunto *Mf* y FAUM sobre el conjunto *Mi*, seleccionando el mismo valor de k en ambos casos. El método de inicialización en el cual se basa esta implementación deriva del mismo método utilizado por la implementación de K-Means de Matlab [26], pero se ha modificado para restringir aún más la aleatoriedad de las semillas iniciales. Esto permite a priori llegar a un buen resultado con una cantidad menor de iteraciones. La función de medición de distancia fue la distancia Euclidiana. Otra diferencia sustancial en la implementación utilizada es el empleo de representación de punto fijo para los valores de los vectores característicos de las muestras. Esto permite incrementar la eficiencia temporal por sobre otras aplicaciones de K-Means en un factor significativo, dependiendo de la cantidad de muestras, características y la precisión en términos de bits de representación de los valores de éstas últimas. En este trabajo este beneficio se presentará en la sección de resultados.

Agrupamiento por FAUM con cantidad fija de clusters

Se aplicó el método FAUM sobre el conjunto *Mi* seleccionando un ajuste manual para establecer la cantidad de clusters igual a 4. Aquí corresponde aclarar algunas particularidades de este método: FAUM en su sentido más amplio consiste en un método determinístico y heurístico que permite descubrir agrupamientos naturales en un conjunto de datos generando a partir de estos datos, histogramas multidimensionales de forma iterativa, estableciendo diferentes tamaños de granularidad (hyper-bines) del histograma y maximizando el equilibrio de la cardinalidad de las clases halladas con dichos hyper-bines. Si bien este método se creó como una solución para descubrir la cantidad natural de grupos en un conjunto de datos, con el objetivo de obtener centroides de grupos y utilizarlos para inicializar otros métodos como K-means++, también es posible utilizarlo para clasificación semi-supervisada si se conoce la cantidad de clases. Como en este trabajo se tiene a priori la clase a la cual pertenece cada muestra aumentada, se utilizaron las muestras sin etiquetar para la ejecución de FAUM y luego se evaluó la eficacia de este método como si fuese un clasificador, utilizando la etiqueta de cada muestra para el cálculo de los valores de la matriz de confusión y de los indicadores estadísticos asociados a esta matriz. En cuanto al detalle del uso, debido a que FAUM trabaja con datos de punto fijo o de tipo entero (para aprovechar las operaciones de desplazamiento de bits con el fin de ganar eficiencia), se utilizó el conjunto *Mi*, el cual se encuentra normalizado y expresado en valores

enteros. De esta manera, se convirtieron los datos iniciales al formato PAM, se verificó su consistencia y se utilizó FAUM acotando el método heurístico para que encuentre soluciones de clusterización con 4 clases.

3 Resultados

En esta sección se presentan los resultados más relevantes de las pruebas que se ejecutaron.

En la Tabla 1 puede observarse cómo se encuentra conformada las muestras utilizadas en las pruebas, luego de culminar la etapa de aumentado de datos. Si bien la cantidad de muestras de las ondas bifásicas, bimodal y normal se han incrementado en factores de 57, 48 y 43 veces respecto de la cantidad original de muestras de esos tipos, en las ondas negativas, por lo mencionado en la sección 2.2, al no haber aún bibliografía suficiente acerca de sus posibles morfologías en el contexto del diagnóstico del Síndrome de Bayés, no se realizó el aumentado y por lo tanto esta clase quedó desbalanceada respecto al resto.

Tabla 1. Muestra relevante

Morfología de la onda P	Cantidad de Muestras
Onda P bifásica	1315
Onda P bimodal	144
Onda P negativa	8
Onda P normal	646
Total	2113

En la Tabla 2 se presentan los resultados de la matriz de confusión especificada para cada tipo de morfología de la onda P y los datos correspondientes para los dos métodos de agrupamiento aplicados. Se puede observar los resultados obtenidos con cada implementación de K-Means++ y FAUM con cantidad fija de clusters.

Tabla 2. Matriz de confusión.

	TP	TN	FP	FN	N
Bifásica	**1315**				**Total**
Kmeans++ Matlab	713	749	49	602	2113
Kmeans++ FAUM	1107	697	101	208	2113
FAUM ajustado	954	696	102	361	2113
Bimodal	**144**				**Total**
Kmeans++ Matlab	144	1928	41	0	2113
Kmeans++ FAUM	144	1891	78	0	2113
FAUM ajustado	144	1360	609	0	2113
Negativa	**8**				**Total**
Kmeans++ Matlab	7	1503	602	1	2113

Kmeans++ FAUM	8	1992	113	0	2113
FAUM ajustado	6	1808	297	2	2113
Positiva	**646**				**Total**
Kmeans++ Matlab	505	1415	52	41	2113
Kmeans++ FAUM	467	1372	95	179	2113
FAUM ajustado	0	1466	1	646	2113

En la Tabla 3 se aprecian los valores de Accuracy, Precision, Recall y f1 Score para la muestra indicada en la Tabla 1.

Tabla 3. Indicadores.

	Acc	**Prec**	**Rec**	**f1 Score**
Bifásica				
Kmeans++ Matlab	0.69	0.94	0.54	0.69
Kmeans++ FAUM	0.85	0.92	0.84	0.88
FAUM ajustado	0.78	0.90	0.73	0.80
Bimodal				
Kmeans++ Matlab	0.98	0.78	1.00	0.88
Kmeans++ FAUM	0.96	0.65	1.00	0.79
FAUM ajustado	0.71	0.19	1.00	0.32
Negativa				
Kmeans++ Matlab	0.71	0.01	0.88	0.02
Kmeans++ FAUM	0.95	0.07	1.00	0.12
FAUM ajustado	0.86	0.02	0.75	0.04
Positiva				
Kmeans++ Matlab	0.91	0.91	0.78	0.84
Kmeans++ FAUM	0.87	0.83	0.72	0.77
FAUM ajustado	0.69	0.00	0.00	0.00

En la Tabla 4 puede observarse un resumen de los indicadores totales obtenidos sobre la matriz de confusión, ponderando cada uno de ellos por la cantidad de muestras de cada clase.

Tabla 4. Indicadores Totales.

	Acc Total	**Prec Total**	**Rec Total**	**f1 Score Total**
K-means++ Matlab	0.78	0.95	0.93	0.94
K-means++ FAUM	0.87	0.87	082	0.84
FAUM con k=4	0.75	0.57	0.52	0.52

En cuanto al tiempo de proceso, es importante destacar que la utilización de la implementación de K-means++ de FAUM, el mismo demoró un tiempo medio de 0.0015 segundos para 2113 muestras, siendo que el tiempo de proceso para 49 muestras es de 0.0009 segundos, lo cual evidencia un factor de aumento de 1.66 siendo que

el factor de aumento de las muestras es de 43.12. De la misma manera, al medir el tiempo de FAUM con k=4, el mismo demoró un tiempo medio de 0.0445 segundos para las 2113 muestras, respecto de un tiempo de 0.005 segundos para las 49 muestras iniciales, lo cual significa un aumento en el factor de tiempo de 8,9 para el mismo factor de aumento de muestras de 43.12. Estas mediciones fueron realizadas con la biblioteca de profiling desarrollada por Fog [38] utilizando en ambos casos el promedio de 25 mediciones y contrastadas con las obtenidas por el comando *time* de la distribución Debian 10 de Linux sobre un equipo Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz con 16 Gb de RAM DDR4.

4 Conclusiones

Debido a los inconvenientes que se presentan con la limitación en la cantidad de señales, se decidió realizar un aumento de datos con las ondas P de los ECGs disponibles. En este trabajo se aplicaron dos métodos de agrupamiento al conjunto de datos, que mediante las técnicas apropiadas, se aumentaron. El objetivo consiste en verificar si es posible agrupar un conjunto de muestras correspondientes a señales de onda P de ECGs para la detección del Síndrome de Bayés.

Observando los resultados de las Tablas 2 y 3 K-means++ FAUM es superior al detectar las ondas bifásicas y negativas. En cambio en la detección de las ondas bimodal y positiva funciona mejor K-means++ de Matlab.

Se puede conluir que en la clase más numerosa (3er grado), todos los métodos funcionaron razonablemente bien, no así en las clases menos numerosas (1er grado, negativa y positiva).

Analizando los valores obtenidos en la 4 es posible concluir que para todas las clases, la implementación de K-Means++ de Matlab logró el mejor resultado considerando el valor de f1-Score.

En líneas generales se ha podido observar que la implementación de K-Means++ de Matlab, analizando los indicadores totales, ha logrado una clasificación mejor que el resto de las soluciones.

Se planifica como trabajo futuro utilizar los datos generados por la técnica de aumentado presentada aquí para: i) mejorar FAUM con cantidad fija de clases modificando sus parámetros de entropía y cardinalidad de modo de obtener mejores resultados, ii) utilizar centroides de referencia para inicializar los clusters de forma predefinida con la idea de utilizar estos algoritmos u otros como clasificadores, iii) aplicar K-Means++ y FAUM sobre los valores resultantes de los componentes derivativos e integrativos de los vectores característicos, entre otras líneas de trabajo.

References

1. Khan, S. S., Ahamed, S., Jannat, M., Shatabda, S., Farid, D. M.: Classifica-tion by clustering (cbc): An approach of classifying big data based on similarities. In Proceedings of International Joint Conference on Computational Intelligence (pp. 593-605). Springer, Singapore (2020).

2. Zhou, Z. H.: A brief introduction to weakly supervised learning. National science review, 5(1), 44-53 (2018).
3. Yu, Z., Kuang, Z., Liu, J., Chen, H., Zhang, J., You, J., Han, G.: Adaptive ensembling of semi-supervised clustering solutions. IEEE Transactions on Knowledge and Data Engineering, 29(8), 1577-1590, (2017).
4. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1476-1485), (2019).
5. Bayés de Luna, A., de Ribot, R. F., Trilla, E., Julia, J., Garcia, J., Sadurni, J., Sagues, F.: Electrocardiographic and vectorcardiographic study of interatrial conduction disturbances with left atrial retrograde activation. Journal of electrocardiology, vol. 18, no 1, pp. 1-13 (1985).
6. Bayés de Luna, A., Cladellas, M., Oter, R., Torner, P., Guindo, J., Marti, V., Iturralde, P.: Interatrial conduction block and retrograde activation of the left atrium and paroxysmal supraventricular tachyarrhythmi. European heart journal, vol. 9, no 10, pp. 1112-1118 (1988).
7. Bacharova, L., Wagner, G. S.: The time for naming the Interatrial Block Syndrome: Bayes Syndrome. Journal of Electrocardiology, vol. 48, no 2, pp. 133-134 (2014).
8. Bayés de Luna, A.: Bloqueo a Nivel Auricular. Rev Esp. Cardiol, vol. 32, no 1, pp. 5-10 (1979).
9. Conde, D., Baranchuk, A.: What a Cardiologist must know about Bayes' Syndrome. Revista Argentina de Cardiología, vol. 82, no 3, pp. 237-239 (2014).
10. Conde, D., Baranchuk, A.: Bloqueo interauricular como sustrato anatómico-eléctrico de arritmias supraventriculares: síndrome de Bayés. Archivos de cardiología de México, vol. 84, no 1, pp. 32-40 (2014).
11. Bayés de Luna, A., Baranchuk, A., Robledo, L. A. E., van Roessel, A. M., artínez-Sellés, M.: Diagnosis of interatrial block. Journal of geriatric cardiology: JGC, vol. 14, no 3, pp. 161 (2017).
12. Baranchuk, A., Torner, P., Bayés de Luna, A.: Bayés Syndrome What Is It? Circulation, vol. 137, no 2, pp. 200-202 (2018).
13. L Kitkungvan, D., Spodick, D. H.: Interatrial block: is it time for more attention? Journal of electrocardiology, vol. 42, no 6, pp. 687-692 (2009).
14. Ariyarajah, V., Puri, P., Apiyasawat, S., Spodick, D. H.: Interatrial block: A novel risk factor for embolic stroke? Annals of Noninvasive Electrocardiology, vol. 12, no 1, pp. 15-20 (2007).
15. Bailey, J. J., Berson, A. S., Garson Jr, A., Horan, L. G., Macfarlane, P. W., Mortara, D. W., Zywietz, C.: Recommendations for Standardization and Specifications in Automated Electrocardiography: Bandwidth and Digital Signal Processing. A report for health professionals by an ad hoc writing group of the Committee on Electrocardiography and Cardiac Electrophysiology of the Council on Clinical Cardiology, American Heart Association. Circulation, vol. 81, no 2, pp. 730-739 (1990).
16. Yochum, M., Renaud, C., Jacquir, S.: Automatic detection of P, QRS and T atterns in 12 leads ECG signal based on CWT. Biomedical Signal Processing and Control, vol. 25, pp. 46-52 (2016).
17. Gritzali, F., Frangakis, G., Papakonstantinou, G.: Detection of the P and T-waves in an ECG. Computers and Biomedical Research, vol. 22, no 1, pp. 83-91 (1989).
18. Lenis, G., Pilia, N., Oesterlein, T., Luik, A., Schmitt, C., Dössel, O.: P wave detection and delineation in the ECG based on the phase free stationary wavelet transformand using in-

tracardiac atrial electrograms as reference. Biomedical Engineering/Biomedizinische Technik, vol. 61, no 1, pp. 37-56 (2016).
19. Gonzalez-Fernandez, R., Rivero-Varona, M., Oca-Colina, G. M.: Detection of P wave in electrocardiogram. En Computing in Cardiology. IEEE, 2013. pp. 515-518 (2013).
20. Chatterjee, H. K., Gupta, R., Mitra, M.: Real time P and T wave detection from ECG using FPGA. Procedia Technology, vol. 4, pp. 840-844 (2012).
21. Ismail, A., Shehab, A., El-Henawy, I. M.: Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations. En Security in Smart Cities: Models, Applications, and Challenges. Springer, Cham. pp. 27-45 (2019).
22. Zavantis, D., Mastora, E., Manis, G.: Robust Automatic Detection of P Wave and T Wave in Electrocardiogram. En 2017 Computing in Cardiology (CinC). IEEE, 2017. pp. 1-4.
23. Franco, L. G., Escobar Robledo, L. A., Bayés de Luna, A., Massa, J. M.: Digitalización de Imágenes de ECG para la Detección del Síndrome de Bayés. En XXIV Congreso Argentino de Ciencias de la Computación, La Plata (2018).
24. Franco, L. G., Escobar Robledo, L. A., Bayés de Luna, A., Massa, J. M.: P-Wave Clustering Methods for Bayès Syndrome Detection. CONAIISI (2020).
25. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, vol. 24, no 7, pp. 881-892 (2002).
26. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. En SODA'07: proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, Philadelphia, PA. 2007, pp. 1027–1035.
27. Berkhin, P.: Survey of clustering data mining techniques. Accrue Software. Inc. TR, San Jose, USA (2002).
28. Curti, H. J., Wainschenker, R. S.: FAUM: Fast Autonomous Unsupervised Multidimensional classification. Information Sciences, vol. 462, pp. 182-203 (2018).
29. Melnykov, V., Melnykov, I., Michael, S.: Semi-supervised model-based clustering with positive ad negative constraints. Advances in data analysis and classification, vol. 10, no 3, pp. 327-349 (2016).
30. Ahn, S., Choi, H., Lim, J., Lee, K. E.: Self-semi-supervised clustering for large scale data with massive null group. Journal of the Korean Statistical Society, vol. 49, no 1, pp. 161-176 (2020).
31. Bayés de Luna, A.: ECGs for beginners. John Wiley & Sons (2014).
32. Polson, N. G., Scott, S. L.: Data augmentation for support vector machines. Bayesian Analysis, 6(1), 1-23 (2011).
33. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., Le, Q. V.: Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 113-123), (2019).
34. Bayés de Luna, A.: Textbook of clinical electrocardiography. Springer Science & Business Media (2012).
35. Murtagh, F., Contreras, P.: Methods of Hierarchical Clustering. CoRR, abs/1105.0121, (2011).
36. Aggarwal, C. C., Reddy, C. K.: Data Clustering: Algorithms and Applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra, (2014).
37. Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier. En Proc. 13th Intl. Conf. Machine Learning. pp. 105-112, (1996).
38. Fog, A.: Pseudo-Random Number Generators for Vector Processors and Mul-ticore Processors. Journal of modern applied statistical methods, 14(1), 23, (2015).