المجلة الدولية للهندسة وتقنية المعلومات
journal homepage:www.ijeit.misuratau.edu.ly

دراسة نقاط البيانات البينية باستخدام خوارزمية (Fuzzy C-Means)

سالم محمد بودبوس
جامعة مصراتة، كلية تقنية المعلومات
s.budabbus@it.misuratau.edu.ly

الملخص — **تلعب طبيعة مجموعة البيانات دورا بارزا في تحديد خوارزمية التجميع المناسبة، ويعتبر تداخل نقاط البيانات أحد أبرز التحديات المؤثرة على أداء الخوارزمية وجودة نتائجها. من جانب آخر، ينظر إلى معيار المسافة على أنه المعيار الوحيد المستخدم لتجميع نقاط البيانات، ويعتبر أحد التحديات المؤثرة التي تحظى بالاهتمام البحثي. في هذه الورقة، تم دراسة النقاط البينية لمجموعة البيانات المتداخلة والمتموضعة بين التجمعات ومدى تأثيرها على أداء خوارزميات التجميع وذلك باستخدام خوارزمية (Fuzzy C-Means) ومعيار درجة العضوية بالإضافة إلى معيار المسافة. أظهرت النتائج أن النقاط البينية لمجموعة البيانات لها تأثير مباشر على أداء خوارزميات التجميع، وخلص البحث إلى أن مجموعة البيانات تحدد طبيعة المعالجة لهذه النقاط، إما باستبعادها وتجاهل تأثيرها أو بوضعها في تجمع مستقل ومن ثم تحليله مع بقية التجمعات.**

الكلمات المفتاحية: **تعلم الآلة، التجميع الصلب، التجميع اللين، المنطق الضبابي، التداخل، Silhouette.**

1. المقدمة

يعتبر مجال تعلم الآلة (machine learning) بكافة فروعه من التقنيات التي لها دور فعال في الاستفادة من البيانات وتوظيفها في بناء نماذج ذكية لوصف البيانات والتنبؤ بالمستقبل.

ينقسم تعلم الآلة إلى فرعين أساسيين هما التعلم بالإشراف (supervised learning) والذي يتطلب معرفة مسبقة بالمخرجات أي أن البيانات معلمة، والتعلم بدون إشراف (unsupervised learning) والذي لا يتطلب معرفة مسبقة بالمخرجات أي أن البيانات غير معلمة.

التجميع (clustering) يعتبر من أهم تقنيات التعلم بدون اشراف، وهو تقنية تعمل على تجميع نقاط البيانات (dataset) في تجمعات بناء على التشابه (similarity) بينها، وهذا التشابه يعتمد على معيار المسافة (distance) وقرب نقاط البيانات من بعضها البعض، كل مجموعة من نقاط البيانات المتشابهة تشكل تجمع واحد، تتمايز فيها عن بقية نقاط البيانات والتي تشكل بدورها أي بقية نقاط البيانات تجمعات أخرى.

تسمى عملية تجميع مجموعة من نقاط البيانات في فئات من النقاط المتشابهة بالتجميع. التجميع عبارة عن مجموعة من نقاط البيانات المتشابهة مع بعضها البعض داخل نفس المجموعة وتختلف عن النقاط الموجودة في المجموعات الأخرى. يمكن التعامل مع مجموعة من نقاط البيانات بشكل جماعي كمجموعة واحدة باعتبارها شكلا من أشكال ضغط البيانات. [17]

نقاط البيانات قد تكون ذات أشكالا محدبة (convex shape) أو أشكالا مقعرة (concave shape)، وقد تكون متداخلة (overlapping) أو غير متداخلة أي متمايزة عن بعضها البعض. وتعتبر نقاط البيانات ذات الأشكال المقعرة والمتداخلة من التحديات الكبيرة لخوارزميات التجميع، وتقييم أدائها يعتمد بشكل كبير على جودة مخرجاتها مع هذا النوع من مجموعة البيانات.

ينقسم التجميع إلى تجميع صعب (hard clustering) وتجميع لين (soft clustering)، التجميع الصعب يعمل على تعيين نقطة البيانات إلى تجمع واحد فقط، أي أن كل نقطة بيانات تنتمي لتجمع واحد فقط، أما التجميع اللين فإن نقطة البيانات قد تنتمي لأكثر من تجمع بدرجة معينة حسب انتمائها للتجمع وعلاقتها ببقية التجمعات، أي أن نقطة البيانات لها درجة عضوية (relationship degree) بنسب مختلفة لأكثر من تجمع.

التجميع اللين يعمل على تعيين درجة عضوية لكل نقطة بيانات، وحسب هذه الدرجة تتدرج نقاط البيانات من نقاط ذات درجة عضوية قوية تتموضع داخل أو بالقرب من مركز التجمع إلى نقاط بيانات ذات درجة عضوية ضعيفة تتموضع على الحدود البينية للتجمعات، وهذه النقاط ذات درجة العضوية الضعيفة يمكن تسميتها بالنقاط الضبابية (Fuzziness points).

في هذا البحث، تم دراسة مجموعة النقاط الضبابية المتموضعة على الحدود البينية للتجمعات، حيث تتميز هذه النقاط بقدرتها على تشكيل نمط مختلف يتمتع بخصائص تختلف عن نقاط البيانات ذات الدرجة العضوية القوية. توفر نقاط البيانات الضبابية رؤى جديدة اعتمادا على معيار الانتماء أو درجة العضوية بدلا من معيار المسافة. ويمكن تحليلها لتوفر معلومات وتفاصيل إضافية للمشكلة محل الدراسة وذلك إما بتجميعها في تجمع جديد يضاف للتجمعات الموجودة مسبقا، أو يمكن حذفها أو تجاهل تأثيرها ثم دراسة وتحليل مجموعة البيانات بشكل مستقل عن النقاط الضبابية للوصول لفهم أعمق وأوضح للمشكلة محل الدراسة.

2. الدراسات السابقة

بناء على الدراسات السابقة، تستخدم خوارزميات التجميع في مجالات مختلفة مثل الرعاية الصحية والتعليم وعلوم الحاسوب والاتصالات والعلوم البيولوجية. يتم استخدام خوارزمية K-Means وخوارزمية Fuzzy C-Means للتجميع وتصنيف المجموعات.

تعتبر خوارزميات التجميع الصلبة مثل (K-Means) مناسبة لمهام التجميع الحصرية. وتعتبر خوارزميات التجميع اللينة مثل (Fuzzy C-Means) مناسبة لمهام التجميع المتداخلة. يُفضل التجميع على أساس قيمة درجة العضوية في المواقف التي تساهم فيها البيانات في مجموعات متعددة [1]. ويعتبر تجميع (Fuzzy C-Means) أكثر دقة من تجميع (K-Means). تسمح خوارزمية (Fuzzy C-Means) بالعضوية التدريجية لنقاط البيانات في المجموعات، بالتالي توفر المرونة للتعبير عن نقاط البيانات التي يمكن أن تنتمي إلى مجموعات متعددة [14].

في مجال الرعاية الصحية، وحسب المرجع [8]، تم استخدام خوارزميات التجميع لتصنيف مجموعات المرضى وتقسيم مرض السكري إلى تجميعات فرعية (subcluster) بناءً على مضاعفات المخاطر والتعريف الجيني والسمات السريرية واختيار العلاج. تشير النتائج إلى أن خوارزميات K-Means وFuzzy C-Means يمكن استخدامها في التصنيف الفرعي لمرض السكري والتنبؤ بمرض السكري.

تم استخدام خوارزمية Fuzzy C-Means في مجال التعليم، لتحديد جودة الخدمة التعليمية عبر معايير الخدمة الأدنى في التعليم. يهدف

استلمت الورقة بالكامل في 28 ابريل 2024 وروجعت في 10 مايو 2024
وقبلت للنشر في 15 مايو 2024
ونشرت ومتاحة على الشبكة العنكبوتية في 08 أغسطس 2024

البحث [9] إلى تطوير نموذج لمجموعات المدارس الابتدائية بناءً على معايير الخدمة الأدنى، وتستخدم خوارزمية Fuzzy C-Means لتجميع المدارس الابتدائية وتقديم نموذج يمكن للمسؤولين استخدامه في تجميع المدارس بناءً على معايير الخدمة الأدنى للتعليم.
كما يوصى المرجع [4] باستخدام خوارزمية K-Means للبيانات الكبيرة والفئوية للحصول على دقة عالية. حيث تم استخدام خوارزميات التجميع في تحليل البيانات وتجزئة بنية البيانات. وتم استخدام خوارزميات ومنهجيات مختلفة للتعامل مع مجموعات البيانات الكبيرة والصغيرة وتجميع البيانات بناءً على الخصائص وأوجه التشابه.
خوارزمية K-Means أسرع في الأداء من خوارزمية Fuzzy C-Means في جميع مجموعات البيانات التي تتمتع بأنماط منتظمة أو غير منتظمة. ومع ذلك، تعتبر Fuzzy C-Means أكثر دقة في تجميع مجموعة بيانات، ولذلك يوصى باستخدام خوارزمية K-Means للبيانات الكبيرة والمتشتتة بسبب سرعة تنفيذها [7].
في المجمل، يستخدم التجميع في علم البيانات لتحليل البيانات وتصنيف المجموعات بناءً على الخصائص المشتركة. يمكن استخدام خوارزميات التجميع مثل K-Means وFuzzy C-Means في العديد من المجالات لفهم البيانات واستخلاص المعلومات الهامة.

3. مشكلة البحث

تأثير نقاط البيانات الضبابية المتموضعة في المنطقة المتداخلة على تجميع البيانات خاصة عند استخدام معيار المسافة فقط، حيث يؤدي وجودها إلى إنتاج تجمعات لا تعبر بشكل دقيق عن المشكلة قيد الدراسة، وإلى فهم غير دقيق لمجموعة البيانات وعلاقاتها وأنماطها.

4. أسئلة البحث

كيف يمكن تجميع مجموعة البيانات باستخدام معيار آخر غير معيار المسافة؟ ما تأثير حذف نقاط البيانات الضبابية من مجموعة البيانات أو فصلها في تجمع جديد على أداء خوارزميات التجميع؟

5. هدف البحث

تحسين أداء خوارزمية (K-Means) من خلال معالجة نقاط البيانات الضبابية إما بحذفها أو فصلها في تجميع جديد وذلك بتطبيق معيار آخر لتجميع البيانات باستخدام المنطق الضبابي (fuzzy logic) ومعيار درجة العضوية.

6. منهجية البحث

ينقسم البحث إلى قسمين لهما علاقة بدراسة نقاط البيانات الضبابية وتأثيرها على أداء خوارزمية التجميع (k-Means)، يبحث القسم الأول تأثير نقاط البيانات ذات درجة العضوية الضعيفة على أداء خوارزمية (k-Means)، ويبحث القسم الثاني مقارنة بين معيار المسافة ومعيار درجة العضوية لتجميع نقاط البيانات ذات درجة العضوية الضعيفة في تجمع مستقل.

تأثير حذف نقاط البيانات ذات درجة العضوية الضعيفة على أداء خوارزمية (k-Means).

1. تطبيق خوارزمية (K-Means).
2. تطبيق خوارزمية (Fuzzy C-Means)، احتساب درجة العضوية لكل نقطة بيانات.
3. حذف نقاط البيانات ذات درجة العضوية الضعيفة.
4. إعادة تطبيق خوارزمية (K-Means).
5. مقارنة أداء خوارزمية (K-Means) قبل وبعد حذف النقاط الضبابية من حيث قيمة معامل (silhouette) لكل نقطة بيانات.

مقارنة معيار المسافة ومعيار درجة العضوية في تجميع نقاط البيانات ذات درجة العضوية الضعيفة في تجمع جديد مستقل.

1. تطبيق خوارزمية (K-Means). عندما (k=3).
2. تطبيق خوارزمية (Fuzzy C-Means)، عندما (c=2)، وتسجيل درجة العضوية لكل نقطة بيانات.
3. تجميع نقاط البيانات الضبابية ذات درجة العضوية الضعيفة في تجميع جديد.
4. المقارنة بين نقاط بيانات التجمع الثالث بمعيار المسافة الناتجة عن خوارزمية (K-Means)، مع نقاط بيانات التجمع الثالث بمعيار درجة العضوية الناتجة عن خوارزمية (Fuzzy C-Means)، وذلك من حيث معامل (silhouette).

7. أدوات البحث

التجميع الصعب يعمل على تعيين نقطة البيانات إلى تجمع واحد فقط، ودرجة عضوية هذه النقطة للتجمع إما تساوي (0) أي لا تنتمي للتجمع أو تساوي (1) أي تنتمي بالكامل لهذا التجمع.
تنتمي خوارزمية (k-means) إلى التجميع المبني على النقطة المركزية (centroid-based clustering) أو يسمى أيضا التجميع المبني على التقسيم (partition-based clustering)، وتأخذ معلمة الإدخال (k)، وتقسم نقاط البيانات (n) إلى تجمعات (k)، بحيث يكون التشابه مرتفعا داخل التجمع (intra-cluster)، ومنخفضا بين التجمعات (inter-cluster). يتم قياس التشابه داخل التجمع بالقيمة المتوسطة (mean value) للنقاط الموجودة داخل التجمع، والتي يمكن اعتبارها المركز الأوسط للتجمع (cluster's centroid). [17].
أما التجميع اللين فإن نقطة البيانات قد تنتمي لأكثر من تجمع بدرجة معينة حسب درجة انتمائها وتماسكها مع نقاط بيانات التجمع، أي أن نقطة البيانات لها درجة عضوية بنسب متفاوتة لأكثر من تجمع تدل على قوة عضويتها فيه. يتمتع التجميع اللين بميزة رئيسية مقارنة بالتجميع الصعب، تشير العضوية لأي نقطة بيانات إلى ما إذا كان هناك تجمع "ثاني أفضل" يكاد يكون بنفس جودة التجمع "الأفضل"، وهي ظاهرة غالبًا ما تكون مخفية عند استخدام تقنيات التجميع الأخرى. [18]
تعتبر خوارزمية (Fuzzy C-Means) تقنية لتجميع البيانات حيث تنتمي كل نقطة بيانات إلى تجمع معين بدرجة معينة تحددها درجة الانتماء. تم تقديم هذه التقنية في الأصل بواسطة جيم بيزديك في عام 1981 كتحسين لطرق التجميع السابقة، الميزة الرئيسية لخوارزمية (Fuzzy C-Means) هي أنها تسمح بالعضوية التدريجية لنقاط البيانات في تجمعات تقاس بالدرجات في النطاق [0،1]. [1].
مقياس المسافة (مقياس القرب) هو مقياس لمساحة الميزة المستخدمة لتحديد مدى تشابه الأنماط [15]. واستنادا إلى المسافة يتم حساب قيمتين لتقييم التجميع، والقيمتان هما: الفصل بين التجمعات (separation)، وتماسك التجمعات (cohesion).
الفصل (separation) يعبر عن مدى تقارب أو تباعد التجمعات عن بعضها البعض (inter-clustering)، والتماسك (cohesion) يعبر عن مدى تقارب أو تباعد النقاط داخل التجمع الواحد (intra-clustering).
يأخذ معامل (silhouette) في الاعتبار المسافات بين التجمعات وداخلها لأي نقطة بيانات معينة. حيث يتم حساب متوسط المسافات (a) لجميع النقاط داخل التجمع لنقطة معينة (x). يتم بعد ذلك حساب المتوسط (b) لجميع التجمعات البينية الأخرى التي لا تحتوي على النقطة (x). يتم بعد ذلك استخدام هاتين القيمتين (a, b) لتقدير معامل (silhouette) للنقطة (x). يصبح متوسط جميع معاملات (silhouette) هو معامل (silhouette) لجميع نقاط مجموعة البيانات [5]. أنظر المعادلة رقم (1)

$$sil = \frac{(b(x) - a(x))}{\max(a(x), b(x))} \quad (1)$$

8. نتائج البحث

لتطبيق فكرة البحث تم استيراد مجموعة بيانات عددية (market_segmentaion) من موقع كاجل (kaggle.com)، تتكون من عدد (2) ميزة، الأولى مستوى رضا العميل، والثانية مستوى ولاء العميل، تم تطبيعها من (1-0)، أنظر شكل رقم (1).

6	1.03	1	0.085	0.915
9	-0.99	1	0.494877	0.505123
10	0.37	1	0.107258	0.892742
9	0.03	1	0.120147	0.879853
3	-1.36	0	0.968866	0.031134
5	0.73	1	0.179129	0.820871

جدول رقم (2): أداء خوارزمية (k-Means)

Relationship Degree	Data Points		Silhouette Score	Silhouette Score	
	C0	C1		C0	C1
nan	23	7	0.481	0.418	0.687
0.50	23	7	0.481	0.418	0.687
0.55	20	7	0.540	0.485	0.697
0.60	19	7	0.565	0.516	0.699
0.65	18	7	0.579	0.530	0.705
0.70	17	7	0.595	0.547	0.711
0.75	13	7	0.663	0.632	0.721
0.80	12	7	0.687	0.663	0.728
0.85	11	7	0.708	0.693	0.733

القسم الثاني: مقارنة معيار المسافة ومعيار درجة العضوية في تجميع نقاط البيانات ذات درجة العضوية الضعيفة في تجمع جديد مستقل.

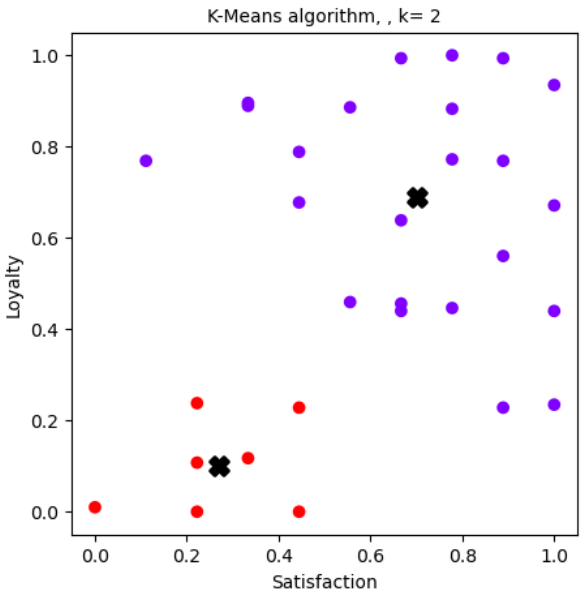
عند البحث عن تجمع بيني بين تجمعين رئيسيين، فإن الخوارزمية الصلبة (k-Means)، أحيانا (حسب طبيعة مجموعة البيانات) لا تعطي تقسيما يظهر بشكل دقيق وواضح التجمع البيني لأنها تعتمد بشكل مباشر على معيار المسافة فقط.

في الشكل رقم (2.1)، يوضح مجموعة بيانات تم تجميعها على (k=2) أي تجمعين، ثم بعد ذلك تم تجميعها على (k=3) أي ثلاث تجمعات، كما في الشكل رقم (2.2)، توضح الأشكال السابقة أن خوارزمية (k-Means) قامت بتجميع مجموعة البيانات إلى ثلاث تجمعات بناء على معيار المسافة. أنظر الجدول رقم (3).

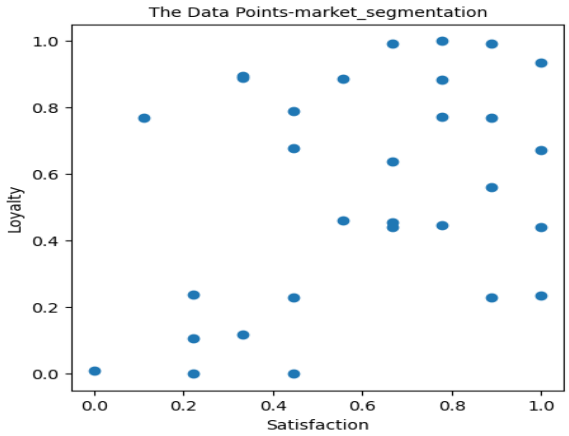
بالنظر إلى الشكل (2.2)، مع تطبيق خوارزمية (k-Means)، التجمع الجديد الثالث الذي بين التجمعين الأول والثاني قد لا يمثل تجمعا حقيقيا، أي لا يمثل تجمعا بينيا بين التجمعين الأول والثاني، أو عدد النقاط داخل هذا التجمع لا يعبر عن التجمع بشكل حقيقي.

جدول رقم (3): تطبيق خوارزمية (k-Means, k=3)

k-means	Data Points			Silh Score	SilhScore		
	C0	C1	C2		C0	C1	C2
3	**14**	**7**	**9**	**0.428**	0.310	0.625	0.45



شكل رقم (2.1)



شكل رقم (1): رسم نقطي لمجموعة البيانات

القسم الأول: تأثير حذف نقاط البيانات ذات درجة العضوية الضعيفة على أداء خوارزمية (k-Means).

جدول رقم (2) يوضح أداء خوارزمية (k-Means) مع مجموعة البيانات بعد حذف نقاط البيانات بناء على درجة العضوية المسجلة في العمود (Relationship Degree)، والتي تم تحديدها مسبقا باستخدام خوارزمية (Fuzzy C-Means)، أنظر جدول رقم (1).

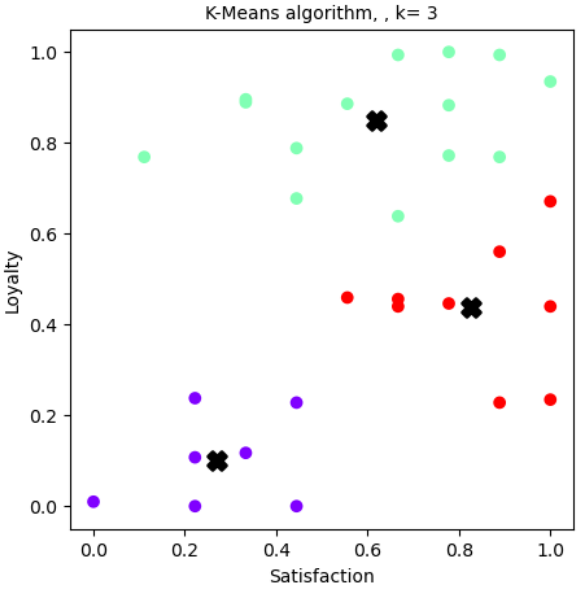
نستنتج من الجدول رقم (2) أن حذف النقاط ذات درجة العضوية الضعيفة يرفع من أداء خوارزمية (k-Means)، ويعمل على التقليل من التداخل (Overlapping) الحاصل بين نقاط البيانات، وذلك حسب قيمة معامل (Silhouette).

حيث تحسّنت قيمة معامل (Silhouette) من (0.481) بدرجة عضوية (0.50) إلى القيمة (0.708) بدرجة عضوية (0.85). وهذا يحقق فصل التجمعات وتمايزها عن بعضها البعض.

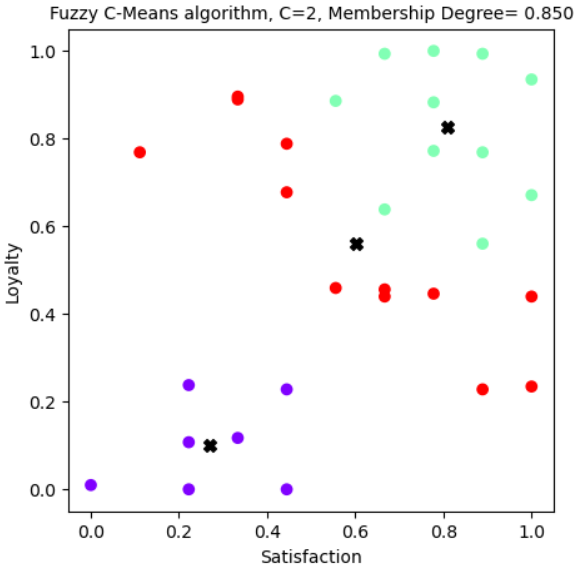
جدول رقم (1): درجات العضوية لنقاط البيانات

Satisfaction	Loyalty	Labels	Deg1	Deg2
4	-1.33	0	0.991503	0.008497
6	-0.28	1	0.477956	0.522044
5	-0.99	0	0.967196	0.032804
7	-0.29	1	0.326425	0.673575
4	1.06	1	0.260339	0.739661
1	-1.66	0	0.876593	0.123407
10	-0.97	1	0.435859	0.564141
8	-0.32	1	0.261536	0.738464
8	1.02	1	0.031965	0.968035
8	0.68	1	0.006515	0.993485
10	-0.34	1	0.252263	0.747737
5	0.39	1	0.248993	0.751007
5	-1.69	0	0.93502	0.06498
2	0.67	1	0.489208	0.510792
7	0.27	1	0.044603	0.955397
9	1.36	1	0.085956	0.914044
8	1.38	1	0.074819	0.925181
7	1.36	1	0.080743	0.919257
7	-0.34	1	0.361579	0.638421
9	0.67	1	0.040745	0.959255
10	1.18	1	0.101301	0.898699
3	-1.69	0	0.942392	0.057608
4	1.04	1	0.26176	0.73824
3	-0.96	0	0.965803	0.034197

أما بالنسبة لخوارزمية (Fuzzy C-Means)، ستعطي (3) تجمعات بناء على معيار المسافة بالإضافة إلى معيار درجة العضوية، مما يعطي تجمعا يمثل الحالة البينية للتجمعين الرئيسيين، أنظر شكل رقم (2.3).
مع تطبيق خوارزمية لينة (Fuzzy C-Means, C=2) سيكون الناتج هو تجمعين فقط، ثم نقوم بحساب النقاط الحدودية الفاصلة بين التجمعين ذات درجة العضوية الضعيفة، بالتالي سيكون لدينا تجمعا ثالثا قد يكون أكثر تمثيلا للنقاط البينية ويعطينا أكثر فهما لنمط البيانات، أنظر الشكل رقم (2.3).



شكل رقم (2.2)

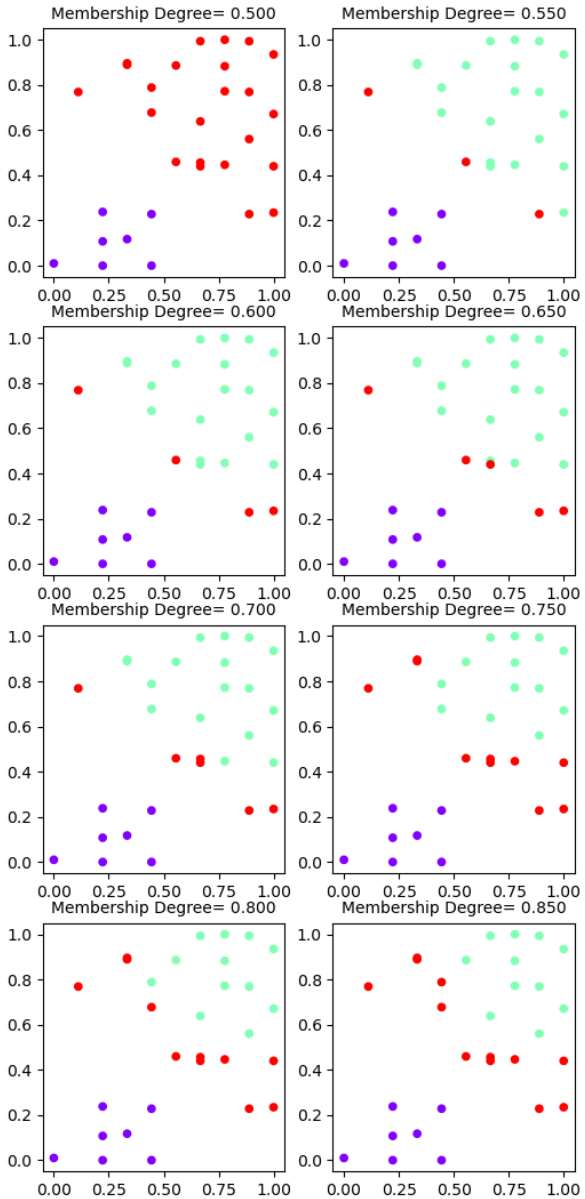


شكل رقم(2.3)

من جانب آخر، يمكن بواسطة قيمة درجة العضوية لنقاط البيانات الناتجة عن خوارزمية (Fuzzy C-Means) التحكم في حجم التجمع البيني، حيث يمكن التدرج في قيمة درجة العضوية حتى الحصول على التجمع البيني المناسب أو العدد المناسب للنقاط البينية للمشكلة قيد الدراسة، أنظر الجدول رقم (4)، يؤدي التحكم في عدد النقاط البينية دورا مهما في جعل التجمعات الأخرى أكثر تميزا عن بعضهما البعض وتقلل من التداخل بينها، أي أن هذه النقاط البينية ستؤدي دورين هما: الأول هو زيادة تمايز التجمعات الأساسية، والثاني هو تجميعها في تجمع مستقل ينتج نمطا جديدا وفهما أكثر لمجموعة البيانات. أنظر شكل رقم (3).

جدول رقم (4): التحكم في حجم التجمع البيني

Degree	Data Points			Silh	SilhScore		
	C0	C1	C2		C0	C1	C2
0.50	23	7	0	0.481	0.418	0.687	nan
0.55	20	7	3	0.270	0.222	0.588	-0.147
0.60	19	7	4	0.317	0.284	0.611	-0.037
0.65	18	7	5	0.304	0.253	0.595	0.079
0.70	17	7	6	0.319	0.252	0.585	0.199
0.75	13	7	10	0.323	0.354	0.627	0.071
0.80	12	7	11	0.329	0.383	0.624	0.084
0.85	11	7	12	0.339	0.424	0.627	0.093



شكل رقم (3): التحكم في عدد النقاط البيني بواسطة درجة العضوية

9. الخلاصة

تم في هذا البحث دراسة النقاط الضبابية المتموضعة بين التجمعات باستخدام خوارزمية (Fuzzy C-Means) ومعيار درجة العضوية. حيث تلعب هذه النقاط دورا مهما في تعزيز التمايز وتقليل التداخل بين التجمعات عندما يتم حذفها أو تجاهل تأثيرها. بالإضافة إلى ذلك، تسهم بشكل مهم في فهم الأنماط والعلاقات في مجموعة البيانات عندما يتم تجميعها في تجمع جديد وتحليلها بشكل مستقل مع التجمعات الأخرى الناتجة عن خوارزميات التجميع.

10.المراجع

[1] Dibya Jyoti Bora, Anil Kumar Gupta, "Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm", International Journal of Computer Trends and Technology (IJCTT), Vol. 10, No. 2, Apr 2014.

[2] Erzhou Zhu, Xue Wang, Feng Liu, "A new cluster validity index for overlapping datasets", IOP Conf. Series: Journal of Physics: Conf. Series 1168, 2019.

[3] Baydaa I. Khaleel, "A Review of Clustering Methods Based on Artificial Intelligent Techniques", Journal of Education and Science (ISSN, 1812-125X), Vol. 31, No. 02, 2022.

[4] Muhammad Faizan, Megat F. Zuhairi, Shahrinaz Ismail, Sara Sultan, "Applications of Clustering Techniques
in Data Mining: A Comparative Study", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 2020.

[5] Said Baadel, "Big Data Analytics: A Tutorial of Some Clustering Techniques", International Journal of Management and Data Analytics (IJMADA), Int. J. Management and Data Analytics, Vol. 1(2), 38-46, 2021.

[6] U. Baid1, S. Talbar, S. Talbar, "Comparative Study of K-means, Gaussian Mixture Model, Fuzzy C-means algorithms for Brain Tumor Segmentation", Conference Paper, DOI: 10.2991/iccasp-16.2017.85, 2017.

[7] Zeynel Cebeci, Figen Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures", Journal of Agricultural Informatics (ISSN 2061-862X), Vol. 6, No. 3:13-23, 2015.

[8] Nashuha Omar, Nisha Nadhira Nazirun, Bhuwaneswaran Vijayam, Asnida Abdul Wahab, Hana Ahmad Bahuri, "Diabetes subtypes classification for personalized health care: A review", Springer Nature B.V. 2022.

[9] Karyo Budi Utomo, Arbain P, and Suminto, "Application of fuzzy C-means algorithm for basic school clustering in Samarinda city based on minimum educational service standard indicators", IOP Conf. Series: Materials Science and Engineering 885, 012002, 2020.

[10] K. Varada Rajkumar, Adimulam Yesubabu, K. Subrahmanyam, "Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset", International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 4, pp. 2760~2770, 2019.

[11] Soumi Ghosh, Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[12] A. B. Raut, G. R. Bamnote, "Soft clustering: An overview", International Journal of Computer and Communication Technology, Vol. 2 Issue 3, 2011.

[13] Sushant Bhargav, Mahesh Pawar, "A Review of Clustering Methods forming Non-Convex Clusters with, Missing and Noisy Data", International Journal of Computer Sciences and Engineering, Volume-4, Issue-3, 2016.

[14] Makhalova Elena, " FUZZY C - MEANS CLUSTERING IN MATLAB", The 7th International Days of Statistics and Economics, Prague, September 19-21, 2013.

[15] Dongkuan Xu, Yingjie Tian, "A Comprehensive Survey of Clustering Algorithms", Springer-Verlag Berlin Heidelberg 2015

[16] A.K. JAIN, M.N. MURTY, P.J. FLYNN, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[17] DM Concepts & Techniques _ Han&Kamber

[18] cluster-analysis_5ed_everitt - EBook