تطوير آليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية[1]

فادي خلوف[2] **أ. د. راكان رزوق[3]**

د. آصف شميس [4]

الملخص

تستخدم تقنية التنقيب في المعطيات بكثرة في إدارة علاقات الزبائن من خلال العديد من المناهج، وتزداد أهمية خوارزميات التنقيب في المعطيات المستخدمة في هذا المجال يوماً بعد يوم. هناك العديد من الأدوات، والطرائق لتحليل المعطيات الخاصة بالزبائن، ويعدُّ التنقيب في المعطيات أحد هذه الأدوات.

يهدف هذا المقال إلى توضيح مفهوم التنقيب في المعطيات، فضلاً عن مفهوم إدارة علاقات الزبائن، وبشكل خاص التقطيع،وتشكيل الزبائن، واستعراض الأدوات، والخوارزميات، ومناهج التنقيب في المعطيات المستخدمة في إدارة علاقات الزبائن، وتطوير نظام يدمج خوارزميتين هما (العنقدة وآلة شعاع الدعم) لمساعدة إحدى شركات التسويق على اتخاذ القرار بشأن عمليات التسويق.

كلمات مفتاحية: التنقيب في المعطيات– آلة شعاع الدعم– العنقدة – التقطيع– تشكيل الزبائن.

[1] أعد البحث في سياق رسالة الدكتوراه للطالب فادي خلوف بإشراف الأستاذ الدكتور راكان رزوق ومشاركة الدكتور آصف شميس.

[2] قسم هندسة الحواسيب والأتمتة– كلية الهندسة الميكانيكية والكهربائية– جامعة دمشق.

[3] قسم هندسة البرمجيات– كلية الهندسة المعلوماتية– جامعة دمشق.

[4] قسم هندسة الحواسيب والأتمتة– كلية الهندسة الميكانيكية والكهربائية– جامعة دمشق.

1- مقدمة:

تستخدم تقنية التنقيب في المعطيات بكثرة في إدارة علاقات الزبائن، وهناك العديد من المناهج، وخوارزميات التنقيب في المعطيات المستخدمة في هذا المجال، لأن إدارة علاقات الزبائن تعدُّ أكثر أهمية في مجال التنافس اليوم. كما توجد العديد من الأدوات، والطرائق لتحليل المعطيات الخاصة بالزبائن، ويعدُّ التنقيب في المعطيات أحدها [1].

تعدُّ إدارة علاقات الزبائن (CRM=Customer Relationship Management) العملية التي تكامل العلاقة بين قطاع الأعمال، وزبائنه. وقد برهن Bhatia [2] بأن إدارة علاقات الزبون هي إستراتيجية العمل المصممة لإرضاء الزبون (customer satisfaction)، وزيادة الدخل، والربحية (profitability). هذا ولا تعدُّ إدارة علاقات الزبائن منتجاً أو خدمةً، وإنما هي إستراتيجية عمل كاملة تُستخدم لتمكن المؤسسات من إدارة علاقاتها مع الزبائن بفعالية.

يُلاحظ أن التسويق على نطاق واسع (mass-marketing) أصبح أضعف من أن يلبي متطلبات الزبائن، بسبب التطورات التي تحدث في قطاع الأعمال اليوم، فضلاً عن التطورات التي حدثت في التقنيات الحاسوبية مما أدى إلى زيادة في معدل استخدام قواعد المعطيات، ودفع قطاع الأعمال إلى تطوير نظام إدارة قاعدة المعطيات الخاصة بعلاقاته بالزبائن، وهذا ما ساعد قطاع الأعمال على تقديم خدمات أكثر فعالية، وفائدة للزبائن [3].

سوف نسعى لتوضيح طريقة استخدام أدوات التنقيب في المعطيات لدعم مفهوم إدارة علاقات الزبائن، وبشكل خاص التقطيع (segmentation)، وتشكيل الزبائن (customer profiling). التقطيع هو التعبير المستخدم لوصف عملية تقسيم الزبائن إلى مجموعات متجانسة على أساس الصفات المشتركة كالعادات مثلاً، أما تشكيل الزبائن فهو وصف للزبائن بحسب صفاتهم مثل العمر، والدخل، والمهنة، ويتم ذلك من خلال بناء نموذج عن سلوك الزبون، وتحديد محددات هذا النموذج، ومن ثَمَّ فإنه يعدُّ طريقة لتطبيق المعطيات الخارجية على جمهور (population) من الزبائن المحتملين، لذلك يمكن استخدامها للتنقيب عن الزبائن الجدد، أو لإزالة الزبائن السيئين (bad customers)، والهدف هو توقع سلوك الزبائن بالاعتماد على معلومات معروفة عن كل زبون [4]، وذلك من خلال توضيح الخطوات المستخدمة لتوقع سلوك الزبائن الجدد من خلال تجميع، وإعداد المعطيات، ومن ثَمَّ إيجاد المستهلك المحتمل، وأهميته في زيادة الأرباح بتطوير نظام برمجي يدمج كلاً من خوارزمية العنقدة، وآلة شعاع الدعم معاً لزيادة الدقة التنبؤية للنموذج المقترح؛ مقارنة بالنتائج التي يتم الحصول عليها في حالة استخدام آلة شعاع الدعم وحدها.

2- إدارة علاقات الزبائن:

تهدف إدارة علاقات الزبائن إلى اكتساب العديد من الزبائن من خلال الاحتفاظ بالزبائن الحاليين، وجذب الزبائن الجدد، واستعادة الزبائن القدماء، وتطوير العلاقات مع الزبائن الحاليين [5]، ومن ثَمَّ يجب على المؤسسات معرفة زبائنها جيداً، والاستجابة لمتطلباتهم بسرعة لتحقيق هذا الهدف، فضلاً عن توقع متطلبات الزبائن قبل إظهار عدم الرضا عن الخدمات التي تقدمها المؤسسة التسويقية [6].

2-1- التقطيع:

التقطيع (segmentation) هي طريقة لإيجاد أكثر اتصال مباشر (targeted communication) بالزبائن، بمعنى إيجاد أفضل طريقة لتقسيم قاعدة الزبائن إلى مجموعات ذات صفات مشتركة كالعمر، أوالجنس، أو الاهتمامات، أو العادات الشرائية، لإدارة علاقات الزبائن من خلال اختيار عينات من الزبائن، بما يمكن المؤسسات من توجيه الخدمات الفضلى لهم، وتصف عملية التقطيع مميزات (characteristics) مجموعات الزبائن (تدعى مقاطع(segments)، أو عناقيد(clusters)) داخل المعطيات، كما يعني التقطيع وضع جمهور (population) داخل مقاطع تبعاً لصفاتهم المتشابهة. مثلاً تتم عملية التقطيع على الزبائن بحسب العمر، والجنس، والدخل [4,7].

2-2- تشكيل الزبائن:

يعدُّ بناء أوصاف (profiles) عن الزبائن خطوة ضرورية لتسويق الخدمات، وإدارة علاقات الزبائن، ومن ثَمَّ تطبيق الاستراتيجيات المختلفة عليهم.

يمكن تقسيم أوصاف الزبائن إلى عدة أنواع:

- الوصف السكاني (demographic profile).
- الوصف السلوكي (behavior profile).
- وغيرها من الأوصاف.

كما يمكن أن تستخدم تقنيات التنقيب في المعطيات لإنشاء أوصاف عن الزبائن من معطيات الزبائن. كمثال يمكن أن تستخدم تقنية أشجار القرار لإنشاء وصف إحصائي للزبائن الأكثر إنفاقاً من خلال المعطيات الإحصائية الخاصة بالزبائن، ومعطيات الإنفاق الأخيرة [8,9].

3–التنقيب في المعطيات :

يهدف التنقيب في المعطيات إلى الحصول على النماذج (patterns) المفيدة، والمهمة، وكذلك استخراج المعرفة من الكميات الكبيرة للمعطيات المتوافرة باستخدام الخوارزميات الإحصائية.

هناك شكلان أساسيان للتنقيب في المعطيات: وصفي (descriptive)، وتنبؤي (predictive). فالتنقيب الوصفي في المعطيات يمكن أن يفسر كعملية للتنقيب في قاعدة المعطيات لاستخراج النماذج المخفية، أو فرضيات للنماذج التي لم تحدد مسبقاً. أما التنقيب التنبؤي في المعطيات فهو عملية استخدام النماذج المكتشفة لتوقع المستقبل من خلال النموذج التنبؤي.

يشبه التنقيب في المعطيات التعليم الآلي الذي يعني التعلم من المعطيات الموجودة لاكتشاف بعض النماذج، وتوقع العينة المستقبلية .

يطور التنقيب في المعطيات عدة نماذج لحل المسائل العملية، وهي: العنقدة (clustering)، وقواعد الارتباط (association rules)، والتصنيف (classification)، والانكفاء (regression)، فالتجميع هو لوضع العناصر المتشابهة داخل المجموعة نفسها، ويمكن أن يستخدم من أجل الوصف، والتنبؤ. في حين تستخدم قواعد الارتباط لوصف النماذج. أمَّا التصنيف، والانكفاء فهما منهجان لإنشاء التوقع [1].

4- استخدام تقنية التنقيب في المعطيات لإدارة علاقات الزبائن :

يمكن للتنقيب في المعطيات عند تطبيقه على مسائل إدارة التسويق (marketing management) أن يحل بفعّالية العديد من المسائل المهمة المتعلقة بالتفاعل مع الزبائن. كما يمكن أن يستخدم التنقيب في المعطيات

من أجل التقطيع، وتشكيل الزبائن، وإدارة المخاطر (risk management)، وغيرها.

لكن التنقيب في المعطيات يمكنه الحصول على النماذج فقط، ولا يستطيع التحقق من صحتها، ولهذا السبب ما يزال التحليل البشري مطلوباً.

يستخدم التنقيب في المعطيات في إدارة علاقات الزبائن في مجالات مختلفة كالبيع بالتجزئة (retailing)، والمصارف (banking)، والتأمين (insurance)، والعناية الصحية (healthcare)، وغيرها من المجالات [1].

5- أدوات و خوارزميات و مناهج و تقنيات التنقيب في المعطيات :

يساعد التنقيب في المعطيات، والأدوات البرمجية على استخراج أوصاف الزبائن من المعطيات بشكل تلقائي. توجد العديد من مناهج التنقيب في المعطيات، ويجري اختيار أحدها ليتم تطبيقه بحسب الهدف من المشروع.

5-1- العنقدة:

هي التقنية التي تضع الكيانات (entities) المتشابهة داخل المجموعة نفسها بالاعتماد على صفات المعطيات المتشابهة في حين توضع الكيانات المختلفة في مجموعات منفصلة. يقاس التشابه (similarity) بواسطة تابع قياس المسافة (distance measure function)، لذلك فإن معنى العناقيد (clusters) يعتمد على تابع المسافة المستخدم.

تشمل تقنيات العنقدة :

- طريقة الجار الأقرب (K-NN=K-Nearest Neighbors أو K-Means).
- الخرائط الذاتية التنظيم (Self-Organizing Maps).
- نوع خاص من الشبكات العصبونية (Artificial Neural Network).

5-1-1- طريقة الجار الأقرب:

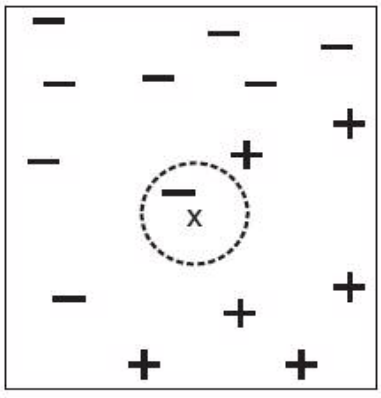
طريقة الجار الأقرب (K-Nearest Neighbor) هي تقنية تنبؤية (predictive) مناسبة لنماذج التصنيف، إذْ تمثل K عدد الحالات المتشابهة، أو عدد العناصر (items) في المجموعة، وتعدُّ معطيات التدريب في طريقة الجار الأقرب هي النموذج فلا يتم بناؤه. عندما يتم تقديم حالة جديدة للنموذج، تبحث الخوارزمية في المعطيات كلّها لإيجاد مجموعة جزئية (subset) من الحالات التي هي أكثر تشابهاً بها، وتستخدمها لتوقع الخرج.

هناك محددان أساسيان في خوارزمية الجار الأقرب :

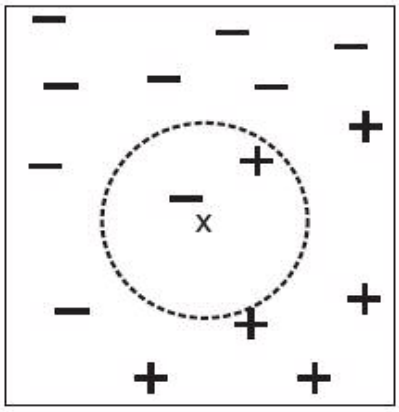
1. عدد الحالات الأقرب ليتم استخدامها (K).
2. وحدة قياس (metric) لقياس التشابه.

يتطلب كل استخدام لخوارزمية الجار الأقرب أن يتم تحديد قيمة موجبة صحيحة لـ k، وهذا يحدد عدد الحالات الموجودة التي يتم البحث عنها عند تقديم حالة جديدة للخوارزمية، كمثال، تشير 4-NN إلى أن الخوارزمية سوف تستخدم الحالات الأربع الأقرب إلى توقع الخرج للحالة الجديدة.

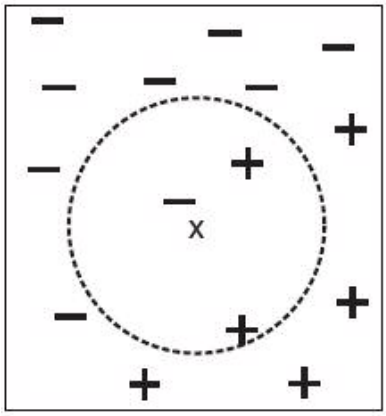
يوضح الشكل (5-1-1) الجار، والجارين، والثلاثة جيران الأقرب لنقطة المعطيات (المثال) المراد تصنيفها، والمتركزة في مركز كل دائرة.



(a) – الجار الأقرب إلى نقطة المعطيات



(b) – الجاران الأقربان إلى نقطة المعطيات



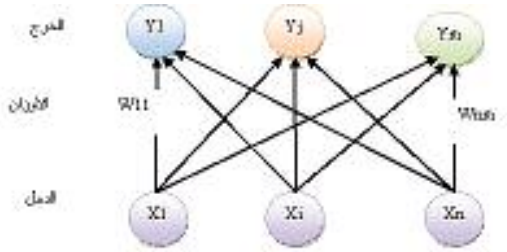
(c)– الثلاثة الجيران الأقرب إلى نقطة المعطيات

الشكل (5-1-1) – الجار، والجاران، والثلاثة جيران الأقرب إلى نقطة المعطيات

إذا كان الجار الأقرب إلى نقطة المعطيات هو المثال السالب كما في الشكل (5-1-1-a) فإن نقطة المعطيات تلك تنتمي للصف السالب، وبحالة كان عدد الجيران هو ثلاثة كما هي الشكل (5-1-1-c)، عندئذٍ يحتوي الجوار (neighborhood) مثالين موجبين، ومثالاً سالباً، وباستخدام مخطط التصويت للأغلبية (majority voting scheme)، فإن نقطة المعطيات تنتمي إلى الصف الموجب. أما في حالة وجود جارين كما في الشكل (5-1-1-b) فيتم اختيار أحد الصفين عشوائياً ليتم تصنيف نقطة المعطيات على أساسه. يتم اختيار k بحيث يتم التغلب على التراكب عندما تكون قيمتها صغيرة جداً بسبب وجود الضجيج في معطيات التدريب، وألا تكون كبيرة جداً فيقوم المصنف بتصنيف الأمثلة (أمثلة الاختبار) بشكل خاطئ [17].

5-1-2- نموذج الخريطة الذاتية التنظيم:

الخريطة الذاتية التنظيم (SOM=Self-Organizing Map) هي أحد أنواع الشبكات العصبونية، وتعتمد على التعلم بوجود معلم (supervised way) لمهمة التجميع، إذْ توجد m وحدة تجميع، مرتبة بمصفوفة من بعد، أو بعدين، وإشارات الدخل هي n صف علاقاتي. في حين يمثل شعاع الوزن لوحدة التجميع كنموذج لعينات الدخل المرتبطة بتلك المجموعة. في أثناء عملية التدريب للخريطة الذاتية التنظيم (SOM)، يجب أن يتم اختيار وحدة التجميع التي تمتلك شعاع الوزن الموافق لعينة الدخل كرابح، لذلك يتم تعديل أوزان الوحدة الرابحة و الوحدات المجاورة لها.



الشكل (5–1–2) – تركيب الخريطة الذاتية التنظيم

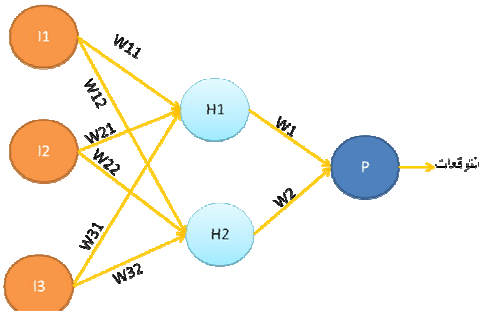
تعدُّ الخريطة الذاتية التنظيم شبكة بسيطة نسبياً – كما هو مبين في الشكل (5-1-2) – مؤلفة من طبقتين للوحدات العصبونية :طبقة الدخل و طبقة الخرج، إذْ تعدُّ صفات الزبائن هي المداخل، والخرج هو تجميع الزبائن.

5-1-3- الشبكات العصبونية الصنعية:

الشبكات العصبونية الصنعية (ANNs=Artificial Neural Networks) هي من بين أكثر خوارزميات الانكفاء، والتصنيف تعقيداً، إذْ تتطلب كثيراً من

المعطيات لتدريبها، ومن ثَـمَّ تـستهلك زمنـاً لهـذا التدريب، ولكن بعد تدريبها يمكنها أن تتوقع الحـالات الجديدة بسرعة كبيرة، حتى في الزمن الحقيقي، كمـا يمكن أن تعطي الشبكات العـصبونية عـدة مخـارج كتوقعات تحدث معاً. لكن الـشبكات العـصبونية لا تعمل إلا على المعطيات الرقمية فقط، ولـذلك يجـب تحويل أية معطيات غير رقمية إلـى أرقـام، فمـثلاً المتحـولات (yes وno)، و(high وlow) يجـب أن تحول إلى 0 أو 1.

تتألف الشبكة العصبونية الأكثر شـيوعاً مـن ثـلاث وحدات : طبقة لوحدات الدخل مرتبطة بطبقـة مـن الوحدات المخفية التي ترتبط بدورها بطبقة لوحـدات الخرج، كما هو مبين في الشكل (5-1-3).



الشكل (5-1-3)- مثال عن شبكة عصبونية أمامية التغذية لأنها تسمح للإشارات بالمرور باتجاه واحد فقط من الدخل إلى الخرج

تُحْسَبُ قيم الخرج للعقد H1، وH2، وعقـدة الخـرج (P)، بالشكل الآتي:

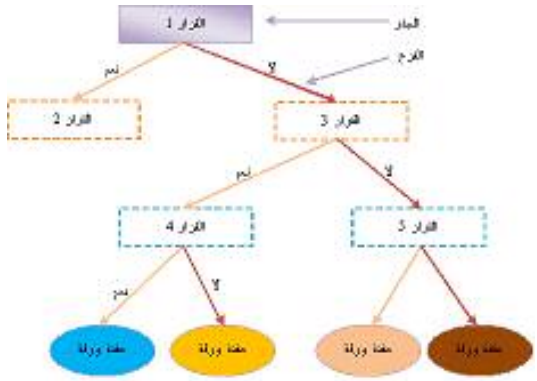
- H1=IF(I1.W11+I2.W21+I3.W31)>0 THEN 1 ELES 0
- H2=IF(I1.W12+I2.W22+I3.W32)>0 THEN 1 ELSE 0
- P=IF(H1.W1+H2.W2)>0 THEN 1 ELES 0

5-2- التصنيف و الانكفاء:

يستخدم التصنيف لتوقع إلى أي مجموعة تنتمي حالـة معطاة. في حين يستخدم الانكفاء لتوقع القيمة لمتحول ذي قيمة مـستمرة (continuous valued variable) بالاعتماد على قيم المتحولات الأخرى. توجد العديـد من تقنيات التصنيف، والانكفاء، وتتـضمن :أشـجار القرار–يمكن استخدام الشبكات العصبونية، وطريقـة الجار الأقرب الموضـحة سـابقاً فـي التـصنيف، والانكفاء.

5-2-1- أشجار القرار:

أشجار القرار (Decision Trees) هي عبـارة عـن تركيب يشبه الشجرة، إذْ تُنَفَّذُ العديد من الاختبـارات للوصول إلى أفضل تسلسل (sequence) مـن أجـل التنبؤ بالمطلوب. بمعنى أن كل اختبار هو إما يـؤدي إلى اختبارات أخرى، أو ينتهي بعقدة ورقـة (leaf)، ويشكل المسار من الجذر إلى عقدة الورقـة الهـدف، القاعدة التي تصنف الهدف (أي التنبـؤ المطلـوب)، وتفسر هذه القاعدة بشكل (IF…THEN) كما هو مبين في الشكل (5-2-1).



الشكل (5-2-1) – تبدأ شجرة القرار من عقدة الجذر، وتُقَسَّمُ المعطيات في كل عقدة لتشكيل فروع جديدة، حتى الوصول إلى العقدة التي لا يمكن تقسيمها إلى أية فروع أخرى (عقدة ورقة).

اجتياز الشجرة من الجذر إلى عقدة الورقة الفضلى تشكل القاعدة التي تصنف الهدف المطلوب

مثال عن قاعدة (IF..THEN) تكون بالشكل الآتي:

- إذا كان القرار1=نعم، ينتهي الاختبار بعقدة ورقية
- إذا كان القرار1=لا، عندئذٍ يتم تطبيـق القـرار3، وهكذا حتى نصل إلى عقدة قرار (مثلاً عقدة القرار4،

أو عقدة القرار5)، وتستمر العملية حتى الوصول إلى عقدة ورقية تحدد التصنيف للحالة الجديدة.

3-5 - تحليل قواعد الارتباط:

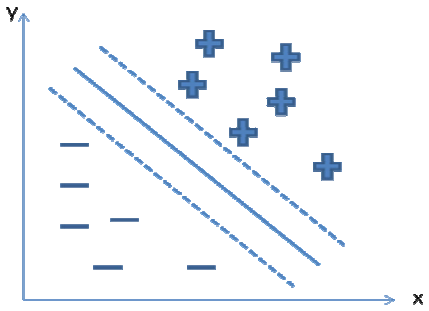
قواعد الارتباط هي أدوات لتحليل المعطيات لاكتشاف القواعد التي تحدد نماذج السلوك، مثلاً ما المنتجات، أو الخدمات التي يميل الزبائن لشرائها معاً؟ كما يمكن استخدام طريقة الارتباط في المصارف لتحليل الحسابات المالية للزبائن، وتحديد مجموعة الخدمات المالية التي يقوم الزبائن بشرائها معاً عادة.

4-5- آلة شعاع الدعم:

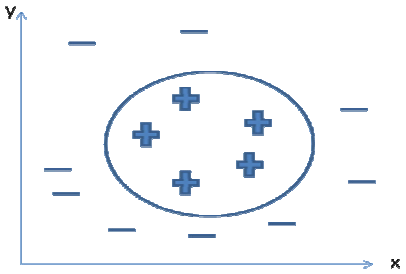
تعدُّ آلة شعاع الدعم (SVM=Support Vector Machine) عبارة عن منهج تدريب ثنائي لتصنيف معطيات التدريب إلى صفين (موجب، وسالب مثلاً) [15]، ويتم تدريبها حسابياً بالاعتماد على مبدأ تقليل الخطأ بين القيم المتوقعة والقيم الفعلية إلى الحد الأدنى، لأنَّ ذلك سيقلل احتمال الخطأ عند تطبيق الخوارزمية على الحالات الجديدة[16].

قام فابنك (Vapnik) [18] بتطوير هذه الخوارزمية لمهمة التصنيف (classification)، والانكفاء (regression) من مجموعة معطيات التدريب. تستطيع آلة شعاع الدعم في مهمة التصنيف إيجاد أفضل سطح فاصل لفصل الأمثلة الموجبة عن الأمثلة السالبة، ويتم اختيار هذا السطح الفاصل بحيث تكون المسافة بين أقرب مثال موجب، وسالب، وهذا السطح الفاصل أكبر ما يمكن، وهذا ما يجعل التصنيف لمعطيات الاختبار صحيحاً [19].

كما تستطيع آلة شعاع الدعم تقسيم المعطيات إلى صفين بواسطة مصنف خطي (linear classifier)، كما في الشكل (5-4-a)، أو بواسطة مصنف غير خطي (nonlinear classifier)، كما في الشكل (5-4-b).



الشكل (a) – المصنف الخطي



الشكل (b) – المصنف غير الخطي

الشكل (5-4) – المصنف الخطي وغير الخطي

6- التطبيق العملي:

بعد أن قُدِّمَ عرض سريع للتقنيات المستخدمة في التنقيب في المعطيات، وخصائص كلٍ منها. نبين فيما يلي طريقة جديدة للتنقيب في المعطيات تدمج تقنيتين هما خوارزمية العنقدة، وخوارزمية آلة شعاع الدعم لحل مسألة التنبؤ بأثر العروض التي تقدمها مؤسسات التسويق في أرباح المؤسسة. تقوم المؤسسات التسويقية بتخزين معلومات عن زبائنها في قواعد معطياتها، بحيث تحتوي تلك المعلومات معلومات شخصية عن كل زبون، مثل السن، والدخل السنوي، والعنوان، وغيرها من المعلومات عن الزبائن، وتستخدمها المؤسسات التسويقية لجذب الزبائن الجدد من خلال استخدام مثلاً المعطيات السكانية المجمعة من عدة مصادر لبناء مجموعات من الزبائن ذوي

صفات شرائية مشتركة، ومنحهم بطاقات التخفيض للزبائن الذين يتوقع أن تزداد مشترياتهم، وكذلك الاحتفاظ بالزبائن الحاليين بتقديم أفضل الخدمات لهم، وتطوير العلاقات ما بين المؤسسة التسويقية، وزبائنها من خلال تحسين جودة الخدمات، والبضائع المعروضة عليهم، مما يؤدي إلى زيادة الأرباح.

تتلخص المسألة المطروحة بأن معظم مؤسسات التسويق تقدم إلى زبائنها بطاقات تسمح لهم بالحصول على المنتجات التي تسوقها هذه المؤسسات بأسعار مخفضة (حسم خاص)، ولكن توزيع هذه البطاقات يجب أن يكون موجهاً بحيث يستفيد منه الزبائن الذين يُتوقع أن تزداد مشترياتهم نتيجة حصولهم على هذه البطاقة، ومن ثَمَّ تزداد أرباح المؤسسة.

سنعرض فيما يلي النتائج التي يمكن الحصول عليها بتطبيق خوارزمية آلة شعاع الدعم بمفردها، وكذلك توضيح نتائج طريقة دمج الخوارزميتين المذكورتين آنفاً للحصول على نتائج أفضل من الحالة الأولى.

6-1-تحضير المعطيات :

تتألف قاعدة المعطيات من 13,500 سجل مستخرجة من قاعدة المعطيات الخاصة بمؤسسة تسويق للمنتجات (أدوات كهربائية- منزلية، وغيرها)، وتحتوي تلك المعطيات معلومات شخصية عن زبائنها كالجنس، والعمر، والدخل السنوي، وتتعلق تلك المعلومات بتصرف الزبائن في السابق (أي المعطيات التاريخية) لإنشاء النموذج الذي يُستخدم لتوقع مستوى الأرباح من الزبائن بعد توزيع بطاقات الحسم، ويُعبر عن ذلك من خلال الواصف AFFINITY_CADR الذي يأخذ إحدى القيمتين : القيمة 0، وتعني أرباحاً منخفضة، والقيمة 1، وتعني أرباحاً مرتفعة، ومن ثَـمَّ تُدعى القيمة المتوقعة المطلوبة لكل زبون (أي 0 أو 1) ليُصَنَّفَ على أساس هذا الواصف.

تخضع معطيات الزبائن الخام إلى المعالجة المسبقة لإزالة المعطيات الضجيجية، والمعطيات المفقودة منها، والمتكررة، قبل استخدام تلك المعطيات كمعطيات دخل للخوارزميتين المذكورتين آنفاً. قُسّمَتْ قاعدة المعطيات المتوافرة إلى ثلاث مجموعات مستقلة بالشكل الآتي: معطيات للتدريب، ومعطيات للاختبار، ومعطيات للتطبيق، بحيث تتألف كل مجموعة من المجموعات الثلاث من 4500 سجل من المعلومات الشخصية.

تستخدم مجموعة التدريب لاستخراج أكبر قدر ممكن من المعلومات (المعرفة) من العينات (أي تدريب النموذج للتمييز بين الزبائن الذين يتم منحهم بطاقات التخفيض، والزبائن الذين لا يتم منحهم إياها)، وتستخدم معطيات الاختبار لتخمين دقة النظام المبني (المدرب). في حين تستخدم مجموعة التطبيق لتخمين الدقة التنبؤية للمصنف على الحالات الجديدة غير المصنفة بعد.

6-2- بناء مستودع المعطيات:

تتألف كل مجموعة من مجموعات المعطيات السابقة من عدة حقول مبينة في الجدول رقم (1) :

الخاصية	المحتوى	نوعها
AGE	العمر	number
ANNUAL_INCOME	الدخل (الراتب)	number
EDUCATION	المؤهل العلمي	Varchar2
MARITAL_STATUS	الحالة العائلية	Varchar2
OCCUPATION	المهنة (الوظيفة)	Varchar2
GENEDER	الجنس	Varchar2
AFFINITY_CARD	بطاقة الحسم (الصفة المطلوبة)	number
ID	معرّف العميل	number
HOUSEHOLD_SIZE	عدد أفراد الأسرة	number

الجدول رقم (1) – بعض الخواص المستخدمة في دراسة عينة الزبائن

3-6- تطبيق خوارزمية آلة شعاع الدعم :

يتم في البداية تطبيق خوارزمية آلة شعاع الدعم على معطيات التدريب لبناء النموذج القادر على التمييز بين الزبائن الذين سوف يتم منحهم بطاقات الحسم عن غيرهم من الزبائن الذين لا يحق منحهم بطاقات التخفيض، وبعد بناء النموذج يتم اختبار دقته بتطبيقه على معطيات الاختبار، فحصلنا على النتائج الآتية:

يبين الجدول (6-3-1) الدقة الكلية عند بناء النموذج باستخدام آلة شعاع الدعم، ومتوسط الدقة عند تطبيقها على معطيات الاختبار، وكذلك الموثوقية التنبؤية (Predictive Confidence) الناتجة عن بناء المصنف المفترض، وتستخدم الموثوقية التنبؤية لقياس مقدار جودة النموذج مقارنة بنموذج التصنيف البسيط (Naïve Classification Model) الذي سوف يتوقع قيم الصف الأكثر حدوثاً بالاعتماد على معطيات البناء.

الدقة الكلية (%)	متوسط الدقة (%)	الموثوقية التنبؤية (%)
78	79.7	59.41

الجدول (6–3–1) – الدقة الكلية، ومتوسط الدقة، والموثوقية التنبؤية الناتجة من تطبيق آلة شعاع الدعم

يشير الجدول (6-3-1) إلى أن النموذج المبني أفضل في بناء نموذج التصنيف للمعطيات 59.41% عن نموذج التصنيف البسيط، كما يبين الجدول (6-3-1) الدقة المتوسطة (Average Accuracy) وهي 79.7% عند تطبيق النموذج على معطيات الاختبار.

يبّين الجدول (6-3-2) ما يُسمى بالتكلفة (Cost)، ويشير إلى مقدار الخطأ الذي يتم الحصول عليه نتيجة التوقع غير الصحيح، وهو مفيد في حالة مقارنة نموذج بآخر، وكلما كانت التكلفة أقل عني هذا بأن النموذج أفضل، ويبين المثال هنا وجود 441 حالة للقيمة الهدف 0، ويتوقعها النموذج بنسبة 75.96% بشكل صحيح، ويتوقع النموذج 83.44% بشكل صحيح لـ 151 حالة للقيمة الهدف 1.

القيمة الهدف	عدد القيم الكلية	الدقة المتوقعة بشكل صحيح (%)
ربح منخفض (0)	441	75.96
ربح مرتفع (1)	151	83.44

الجدول (6–3–2)– التكلفة

يبين الجدول (6-3-3) ما يسمى بمصفوفة الشك التي تستخدم لحساب عدد التوقعات الصحيحة، والخاطئة الناتجة عن النموذج المبني مقارنة بالتصنيفات الفعلية لمعطيات الاختبار.

عدد الحالات	الربح المنخفض المتوقع (0)	الربح المرتفع المتوقع (1)
الربح المنخفض الفعلي (0)	335	106
الربح المرتفع الفعلي (1)	25	126

الجدول (6–3–3) – مصفوفة الشك

تعرض مصفوفة الشك عدد التوقعات الصحيحة، والخاطئة الناتجة عن النموذج المبني مقارنة بالتصنيفات الفعلية في معطيات الاختبار، وهي عبارة عن مصفوفة $n \times n$ صف، إذْ إنَّ n هو عدد الصفوف.

يبّين الجدول (6-3-3) مثالاً عن مصفوفة الشك من أجل نموذج التصنيف الثنائي، وتمثل السطورعدد التصنيفات الفعلية في معطيات الاختبار، وتمثل الأعمدة عدد التصنيفات المتوقعة الناتجة عن النموذج المبني. يُشير هذا النموذج إلى توقع 25 مرة للقيمة 0 (false negative) عندما القيمة الفعلية هي 1، وتدعى بالتوقعات السالبة الخاطئة (false-negative predictions). في حين يُشير توقع 106 مرة للقيمة 1 (false positive) عندما القيمة الفعلية هي 0 وتدعى بالتوقعات الموجبة الخاطئة (false-positive predictions).

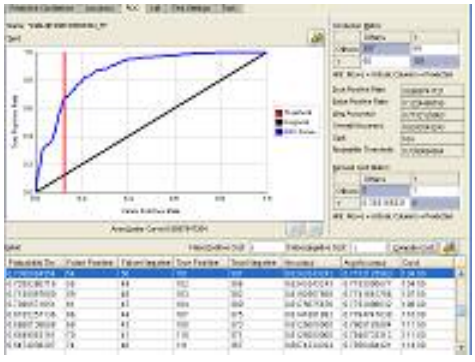
يُستخدم **منحنى مميزات تشغيل المستقبل** (ROC=Receiver Operating Characteristic)

تجريب التحليل ("what-if")، وهو يعدُّ مقياساً آخرَ من أجل مقارنة القيم الهدف المتوقعة، والفعلية في نموذج التصنيف، حيث يطبق التصنيف الثنائي، ويتطلب التعيين للصف الموجب.

يمكن استخدام ROC لكسب فكرة عن مقدرة اتخاذ القرار للنموذج، أي كيف من المحتمل أن يتوقع النموذج الصف الموجب، أو السالب بشكل صحيح ؟

يقيس ROC تأثير التغيرات في العتبة الاحتمالية (probability threshold)، وهو نقطة القرار المستخدمة بواسطة النموذج من أجل التصنيف، فالعتبة الاحتمالية الافتراضية للتصنيف الثنائي هو 0.5، وعندما يكون الاحتمال للتوقع هو 50% أو أكثر، فإن النموذج يتوقع ذلك الصف، وبحالة كان الاحتمال أقل من 50%، فإنه يتم توقع صف آخر.

المنحنى ROC عبارة عن مخطط ذي محورين x-y، يمثل فيه المحور x نسبة الموجب الخاطئة (false positive rate)، أو يسمى معدل التحذير الخاطئ، ويمثل المحور y فيه نسبة الموجب الصحيحة (true positive rate)، أو يسمى بمعدل الاكتشاف، كما هو مبين في الشكل (6-3). تمثل الزاوية العليا اليسارية الموقع المثالي على مخطط ROC، مشيرة إلى نسبة الموجب الصحيحة العالية ونسبة الموجب الخاطئة الدنيا.

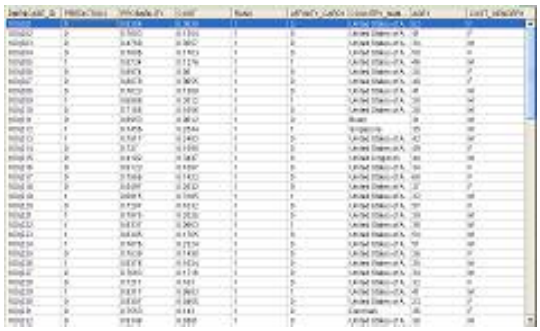


الشكل (6-3) – منحنى مميزات تشغيل المستقبل

يبّين الشكل (6-3) أنَّ العتبة الاحتمالية (probability threshold) وهي الاحتمال الأدنى للصف الموجب المتوقع كانت 0.739008 مع توقع 101 مرة للحالات الموجبة بشكل صحيح،، مقابل 387 مرة لتوقع الحالات السالبة بشكل صحيح.

تقيس المنطقة تحت المنحنى (AUC=Area Under the Curve) المقدرة المميزة لنموذج التصنيف الثنائي، فكلما كانت هذه المنطقة أكبر كان الاحتمال أكبر بأن الحالة الموجبة الفعلية سوف تكون مسندة إلى احتمال أكبر بأن تكون موجبة مقارنة بالحالة السالبة الفعلية، ويعدُّ القياس AUC مفيداً بشكل خاص من أجل مجموعة المعطيات ذات توزيع الهدف غير المتوازن (أي أن صف واحد يهيمن على بقية الصفوف).

يّبين الشكل (6-3-1) نتيجة تطبيق آلة شعاع الدعم على معطيات جديدة من الزبائن، وذلك ليتم تصنيف الزبائن الجدد إلى عملاء يتم منحهم بطاقات الحسم الخاصة بالشركة التسويقية أم لا.



الشكل (6-3-1) – تطبيق آلة شعاع الدعم لتصنيف المعطيات للزبائن الجدد

يبّين الشكل (6-3-1) مثلاً بأن العميل ذا المعرّف 100.002 لا يُمْنَحُ بطاقة الحسم الخاصة بشركة التسويق باحتمال 76.83% وبتكلفة 13.54%، حيث تستخدم الخوارزمية الحقل (Prediction) الذي يأخذ إحدى القيمتين :القيمة 0 لا يمنح بطاقة الحسم، أو القيمة 1 يُمْنَحُ بطاقة الحسم، ويتم التأكيد على ذلك

بالحقل (PROBABILITY) الذي يبيّن في هذه الحالة احتمال 76.83% بأن هذا العميل سوف لن يمنح بطاقة الحسم، وهذا مبين من خلال الحقل (AFFINITY_CARD) الذي يأخذ القيمة 0 للدالة على عدم منح كل الزبائن الذين أعمارهم 62 عاماً، وجنسهم أنثى أية بطاقة حسم لشراء منتجات الشركة التسويقية بأسعار مخفضة.

6-4- دمج خوارزمية العنقدة وآلة شعاع الدعم :

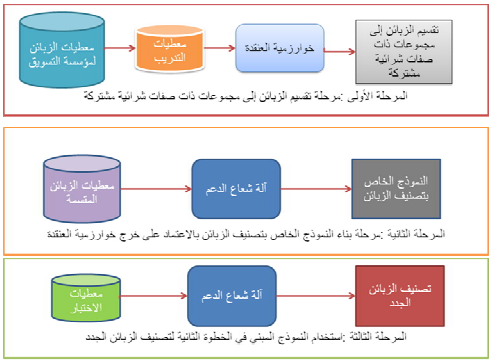
يقتضي النموذج المقترح دمج كلٍّ من خوارزميتي العنقدة، وخوارزمية آلة شعاع الدعم، وذلك للتغلب على المشكلة التي تعاني منها آلة شعاع الدعم عند وجود عدد كبير من أشعة الدعم، حيث تقوم خوارزمية العنقدة بتقليل عدد أشعة الدعم، كما هو مبين في الشكل (6-4-1) الذي يبيّن المخطط الصندوقي للمنهج المقترح.

يوضح الشكل (6-4-1) الخطوات العملية الرئيسية المستخدمة في تطوير هذا النظام :

1. المرحلة الأولى :تجميع معطيات الزبائن الخاصة بمؤسسة التسويق التي تتم دراستها، وتقسيمها إلى معطيات للتدريب، وتستخدم لتقسيم الزبائن إلى مجموعات ذات صفات شرائية مشتركة، باستخدام خوارزمية العنقدة لإيجاد الزبائن المتشابهين داخل الحالات التاريخية، ومن ثَمَّ تقليل عدد أشعة الدعم المستخدمة في آلة شعاع الدعم لبناء نموذج التصنيف التي تعدُّ مشكلة لآلة شعاع الدعم عندما يكون عدد أشعة الدعم كبيراً جداً، مما يؤدي بالنتيجة إلى زمن كبير لبناء نموذج التصنيف.

2. المرحلة الثانية: تُستَخْدَمُ خوارزمية آلة شعاع الدعم لبناء النموذج الذي سوف يستخدم في المرحلة اللاحقة لتصنيف الزبائن الجدد.

3. المرحلة الثالثة: يُستَخْدَمُ النموذج المبني سابقاً على معطيات الاختبار التي تستخدم لاختبار دقة النموذج في تصنيف الزبائن الجدد.



الشكل (6-4-1)-خطوات تصميم النموذج المقترح

اِسْتُخْدِمَتْ 10 حالات للجار الأقرب بسبب أنها حققت أفضل دقة تنبؤية عند استخدامها مع SVM كما هو موضح في الجدول (6-4-1)، حيث يتم تحديد عدد حالات الجار الأقرب، وتطبيق خوارزمية K-means على معطيات التدريب، قبل استخدام خوارزمية SVM.

عدد الحالات للجار الأقرب	الدقة التنبؤية بعد تطبيق SVM
4	80.8%
5	80%
6	78.7%
7	74.9%
8	75.7%
9	78.3%
10	81.35%

الجدول(6-4-1)-الدقة التنبؤية بحسب عدد حالات الجار الأقرب

تُستَخْدَمُ النتائج التي يتم الحصول عليها من تطبيق خوارزمية العنقدة لبناء نموذج التصنيف باستخدام آلة شعاع الدعم لنحصل على النتائج الآتية:

لوحظ تحسن ملحوظ في الموثوقية التنبؤية المحققة عند دمج خوارزمية العنقدة، وخوارزمية آلة شعاع الدعم معاً، إذْ حصلنا على موثوقية تنبؤية 100% مقابل موثوقية تنبؤية 59.41% عند تطبيق آلة شعاع الدعم وحدها مما يشير إلى تحسن في الموثوقية

التنبؤية للنموذج المبني، كما تشير النتائج إلى أن دقة النموذج المبني هي 100% عند تطبيق النموذج على معطيات الاختبار.

يّبين الجدول (6-4-1) ما يسمى بالتكلفة، ويشير إلى مقدار الخطأ الذي يتم الحصول عليه نتيجة التوقع غير الصحيح.

القيمة الهدف	عدد القيم الكلية	الدقة المتوقعة بشكل صحيح (%)
ربح منخفض (0)	442	100
ربح مرتفع (1)	146	100

الشكل (6–4–1) –التكلفة

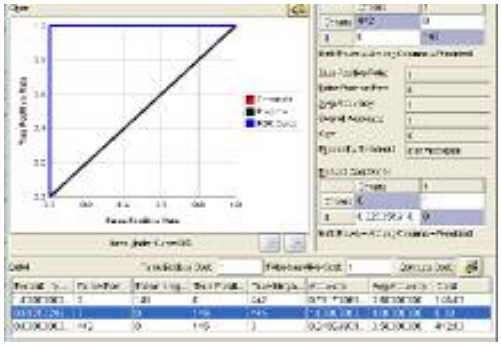
يشير الجدول هنا إلى وجود 442 حالة للقيمة الهدف 0 ويتوقعها النموذج بنسبة 100% بشكل صحيح،، ويتوقع النموذج 100% بشكل صحيح لـ 146 حالة للقيمة الهدف 1.

كما يُظهر الجدول (6-4-2) مصفوفة الشك، ويُشير هذا النموذج إلى توقع 0 مرة للقيمة 0 (false negative) عندما القيمة الفعلية هي 1 وتدعى بالتوقعات السالبة الخاطئة (false-negative predictions). في حين يُشير توقع 0 مرة إلى القيمة 1 (false positive) عندما القيمة الفعلية هي 0 وتدعى بالتوقعات الموجبة الخاطئة (false-positive predictions).

عدد الحالات	الربح المنخفض المتوقع (0)	الربح المرتفع المتوقع (1)
الربح المنخفض الفعلي (0)	442	0
الربح المرتفع الفعلي (1)	0	146

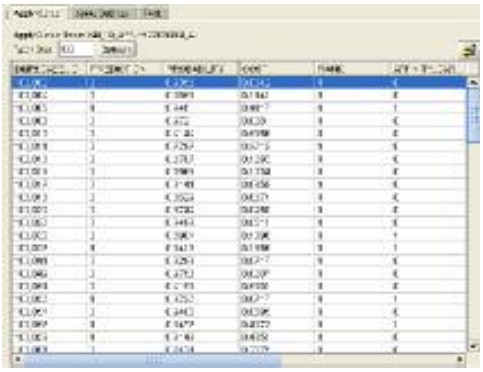
الجدول (6–4–2) – مصفوفة الشك

بينما يبّين الشكل (6-4-2) منحنى مميزات تشغيل المستقبل للنموذج المقترح، ويبّين الشكل بأن المنطقة أسفل المنحى هي أكبر منها الموجودة في الشكل (6-4-3) لآلة شعاع الدعم وحدها.



الشكل (6–4–2) – منحنى مميزات تشغيل المستقبل للنموذج المقترح

يبّين الشكل (6-4-3) تطبيق المنهج المقترح على مجموعة من معطيات الزبائن الجدد ليصنفوا بناءً على الخوارزمية المقترحة.



الشكل (6–4–3)–نتيجة تطبيق دمج خوارزمية العنقدة، وخوارزمية آلة شعاع الدعم على معطيات التطبيق الجديدة.

يُشير الشكل (6-4-3) إلى أنّ الزبون ذا معرّف 100.002 لن يُمْنَحَ بطاقة الحسم الخاصة بشركة التسويق، كما بّينا سابقاً عند تطبيق آلة شعاع الدعم، ولكن الاختلاف هنا بأن الاحتمال هو 93.59% والتكلفة هي 6.42% مقابل 76.83% وتكلفة 13.54%، ونظراً إلى أنَّ التكلفة منخفضة ناتجة عن الخوارزمية المطورة فهي تعدُّ أفضل من حالة استخدام خوارزمية آلة شعاع الدعم وحدها، كما توجد بعض الحالات الشاذة مثل العميل ذي المعرّف 100.032 فسوف تعدُّ في هذه الحالة آلة شعاع الدعم أفضل من دمج الخوارزميتين معاً.

6-5- المقارنة النهائية:

بمقارنة النتائج من مقارنة تطبيق آلة شعاع الدعم، ونتائج المنهج المقترح تبين تحسن في الموثوقية التنبؤية التي كانت 59.41% لدى تطبيق خوارزمية آلة شعاع الدعم وحدها، مقابل 100% لدى تطبيق المنهج المقترح من تطبيق الخوارزميتين (العنقدة، وآلة شعاع الدعم)، وبمقارنة الجدولين (6-3-3) و(6-4-1) لوحظ تحسن النموذج بتوقع القيم السالبة الخاطئة التي هي هنا 0 مرة مقابل 25 مرة. أمَّا توقع القيم الموجبة الخاطئة فهي 0 مرة مقابل 106 مرة فكانت عند تطبيق آلة شعاع الدعم وحدها.

كما يمكن استخدام الشك المبينة في الجدولين المشار إليهما سابقاً لاستخراج معلومات مهمة أخرى تفيد في مقارنة النموذجين المقترحين، ويبّين الجدول (6-5-1) توزيعات الأعمدة، والصفوف من أجل التصنيف الثنائي.

التصنيف الحقيقي	التصنيف المتوقع به	
الحالات	0	1
0	TN (True Negative)	FP (False Positive)
1	FN (False Negative)	TP (True Postive)

الجدول (6–5–1)– توزيعات الأعمدة والصفوف للتصنيف الثنائي

يمكن أن يتم استخراج المعلومات التالية من هذه المصفوفة :

$$TN_{rate} = \frac{TN}{FP + TN}$$

$$FP_{rate} = \frac{FP}{FP + TN}$$

$$FN_{rate} = \frac{FN}{FN + TP}$$

$$TP_{rate} = \frac{TP}{TP + FN}$$

كما يمكن حساب معدل الخطأ، والدقة من خلال العلاقات الآتية:

$$Error_{rate} = \frac{FP + FN}{TN + FP + FN + TP}$$

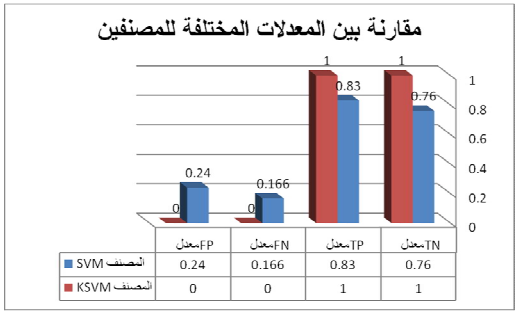
$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

يبّين الجدول (6-5-2) تخمين المصنف لكلٍ من خوارزمية آلة شعاع الدعم (SVM)، والمصنف المقترح من دمج خوارزمية العنقدة، وآلة شعاع الدعم (KSVM).

المصنف	نسبة الخطأ (%)	TN	FP	FN	TP	معدل TN	معدل TP	معدل FN	معدل FP	الدقة (%)
SVM	22	335	106	25	126	0.76	0.83	0.166	0.24	78
KSVM	0	442	0	0	146	1	1	0	0	100

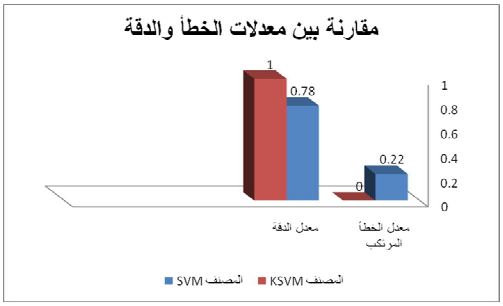
الجدول (6–5–2) – تخمينات المصنفات المقترح المقارنة بينها

عند عرض النتائج بشكل بياني للمعدلات المختلفة المبينة في الجدول (6-5-2) نتيجة تطبيق خوارزمية آلة شعاع الدعم، والدمج الناتج عن خوارزمية العنقدة، وخوارزمية آلة شعاع الدعم :



تشير النتائج إلى أن المصنف SVM حقق معدل TP هو 0.83% مقابل 100% للمصنف المقترح؛ مما يعني بأن المصنف المقترح سوف يتمكن من تحديد الزبائن الذين سوف يمنحون بطاقات الحسم بنسبة أكبر من نموذج المصنف SVM وحده. بينما حقق النموذج المقترح KSVM معدلات FN، و FP، و TN هي :

0%، و 0%، و 100% على التوالي، مقابل معدلات لنموذج SVM هي : 16.6%، و 24%، و 76%.



يشير المخطط البياني إلى أنَّ دقة التنبؤ للمصنف KSVM هي 100% مقابل 78% لمصنف SVM، أمَّا معدل الخطأ الذي يرتكبه المصنف فكان 0% لمصنف KSVM مقابل 22% لمصنف SVM.

7- الخاتمة:

قدمت هذه المقالة توضيحاً عن مفهوم إدارة علاقات الزبائن، وبشكل خاص التقطيع، وتشكيل الزبائن، وطريقة استخدام أدوات التنقيب في المعطيات لدعم عملية اتخاذ القرارات، من خلال توضيح الخطوات المستخدمة لتوقع سلوك الزبائن من خلال تجميع المعطيات وإعدادها، ومن ثَمَّ إيجاد الزبون المحتمل، وأهمية هذا العميل في زيادة الأرباح.

كما تم عرض تطوير نظام يستخدم كلاً من خوارزمية العنقدة K-Means و آلة شعاع الدعم (SVM) معاً، وذلك للتغلب على مشكلة العدد الكبير لأشعة الدعم التي تعاني منها الآلة شعاع الدعم، وقد أظهرت النتائج بأن دمج الخوارزميتين معاً أدى إلى تحسين دقة النموذج المقترح بنسبة 81.35% في تصنيف الزبائن الجدد الذين قد يزيدون من إنفاقهم بحالة مُنحُوا بطاقات الحسم بالمخزن، مقابل 79.7% عند استخدام الآلة شعاع الدعم وحدها.

توجد بعض المقترحات التي يمكن تجربتها لرؤية مدى تحسن الدقة التنبؤية للنموذج المقترح، وهي :

- استخدام عدد عناقيد أقل مما هو مقترح في هذه المقالة، ورؤية مدى تحسن الدقة التنبؤية للنموذج المقترح مقارنة بآلة شعاع الدعم.
- دمج تقنيات أخرى من تقنيات التنقيب في المعطيات المذكورة سابقاً في هذه المقالة، ومقارنة النتائج.

8- مصطلحات علمية:

المصطلح الأجنبي	المقابل العربي
Customer Relationship Management	إدارة علاقات الزبائن
Data Warehouse	مستودع المعطيات
Profitability	الربحية
Decision Trees	أشجار القرار
Clustering	العنقدة
Prediction	التوقع
Classification	التصنيف
Association Rules	قواعد الارتباط
Artificial Neural Networks	الشبكات العصبونية الصنعية
Data Mining	التنقيب في المعطيات
Risk Management	إدارة الخطر
segmentation	التقطيع
Customer Profiling	تشكيل الزبائن
Case Based Reasoning	محاكاة معتمدة على الحالة
Self Organizing Map	الخريطة الذاتية التنظيم
Profile	وصف

9- المراجع :

1- Junling Gao, Xiaojin Dong, data mining in customer relationship management ,2003.

2- Anil, Bhatia, Customer Relationship Management, 1st ed., toolbox Portal for CRM, 1999.

3- Chang Pei-Chann, Wang Yen- Wen, apply case based marketing strategy in customer relationship management system,2004.

4- Catherine Bounsaythip and Esa Rinta-Runsala ,overview of data mining for customer behavior modeling , 2001.

5- Roger Ahern, "Sitting on a data goldmine", http://www.hnc.com/pdf/data_goldmine.pdf .

6- Catherine Bounsaythip and Esa Rinta-Runsala, "Overview of data mining for customer behavior modeling",

http://www.vtt.fi/tte/datamining/publications/customerprofiling.pdf
7- P. Holmes, customer profiling and modeling , in DMG Direct, Direct marketing association , http://www.dirmarketing.com/dmginc/page15.html.
8- Graham Williams , data mining desktop survival guide , 2004.
9- SAS customer segmentation , http://www.sas.com/solutions/crm/segment/index.html, 2007.
10- Ruiyuan Guo and Ajith Abraham , analyzing call center performance : a data mining approach, 2000.
11- Stacey L. Schertel , data mining its potential use in textiles , 2002.
12- Jasmin Catovic , Frank Moritz , Yue Sun , Andre Bunger , classification of credit applications using data mining , 2002 .
13- Vasilis Aggelis , Dimitris Christodqulakis, data mining for decision support in e-banking area , 2003.
14- data mining, www.postech.ac.kr/topic/dm_e.html.
15-Valdimir Vapnik, The Nature of statistical learning theory, Springer, New York, 1995.
16-Ruiyuan Guo, Ajith Abraham, Marcin Paprzycki, Analyzing call center performance: a data mining approach, 1999.
17-S. Cost, S. Salzberg, A weighted nearest neighbor algorithm for learning with symbolic features, Machine Learning, 1993.
18-Vapnik, V. N. ,The nature of statiscal learning theory, Springer, 1995.
19-http://www.kernel-machines.org.

تاريخ ورود البحث إلى مجلة جامعة دمشق: 2009/10/4.