تطوير منهجية جديدة لبناء شجرة القرار باستخدام طريقة هجينة تعتمد على الخوارزمية الجينية وخوارزمية النمل

الطالبة: هدى علي حبش

جامعة البعث – كلية العلوم – قسم الرياضيات

المشرف المشارك: د. عبد الله العمر المشرف: د. زكريا زكريا

الملخص

تُعدّ نظم دعم اتخاذ القرار ناتجاً لتطوُّر تقنية المعلومات بشكل كبير، وتركِّز هذه النظم على توفير الدعم المناسب لتحسين جودة القرار والتي تعتمد على توفُّر المعلومات الكافية وملاءمة النماذج لتحليل المشكلة.

سندرس في هذا البحث خوارزميات إنشاء أشجار القرار وتطويرها، عن طريق اقتراح خوارزمية هجينة تعتمد على دمج الخوارزمية الجينية مع خوارزمية النمل لبناء شجرة القرار باستخدام منهجية جديدة تستفيد من خصائص الخوارزميتين معاً، وهما: ميزة وراثة الصفات الجيدة في الخوارزميات الجينية وميزة التواصل المجتمعي في خوارزمية النمل بهدف الوصول إلى شجرة القرار المثالية التي تتفوَّق على المنهجيات المتبعة سابقاً، ويعرض تفاصيل هذه المنهجية ويوضح نتائجها على مجموعة بيانات حقيقية حيث تظهر المقارنة تحسُّن نتائج هذه الطريقة على نتائج الطرق السابقة في مجال بناء أشجار القرار.

الكلمات المفتاحية:

نظم دعم اتخاذ القرار، أشجار القرار، الخوارزميات الجينية، خوارزمية النمل، الوراثة، التواصل المجتمعي.

Developing a New Methodology to Construct Decision Trees by Using a Hybrid Method Based on Genetic Algorithms and Ant Colony Optimization

Huda Ali Habash
ALBaath University – College of Science – Mathematics department

Abstract:

Decision support systems are considered the result of the advanced development of information technology, as these systems focuses on providing suitable support to improve decision quality that depends on information sufficiency and model fitness to analyze the problem. This research aims at studying and developing decision tree construction algorithms, it suggests a hybrid algorithm that depends on combining both genetic algorithm and ant colony optimization to build a decision tree using a new methodology that makes use of both approaches, which are: the ability to develop good individuals in genetic algorithms and the social communication features of ant colonies. To achieve an optimal decision tree that can overcome available methodologies. This research shows the details of this methodology and elaborates its results by testing it on a real life dataset, the results confirms the superiority of this method compared to previous ones in the field of decision tree construction.

Key Words:

Decision support systems, Decision trees, Genetic algorithm, Ant colony optimization algorithm, Genetics, Community communication.

1. مقدمة:

في السنوات الأخيرة أصبحت هناك حاجة مُلحة لاستخراج المعلومات من القيم غير المؤكدة وغير الدقيقة والغامضة بالتوازي مع القيم الموجودة في قواعد البيانات.

تعتبر قواعد البيانات اليوم من المجالات الهامة والمستخدمة بشكل واسع في هذا العصر.. عصر المعلوماتية. ومع ازدياد قواعد البيانات وضخامة محتوياتها كان لابد من ظهور أدوات وخوارزميات تساعد في التنقيب في البيانات وإعطاء المستخدم المعلومات المفيدة منها. ونتيجة لذلك ظهر مجال هام من مجالات الذكاء الصنعي يدعى التنقيب في البيانات (Data Mining) كتقنية تهدف إلى استخراج المعرفة من كميات هائلة من البيانات، وهي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية، والتنقيب في البيانات هو عملية اكتشاف الارتباطات والأنماط والاتجاهات الجديدة المفيدة من خلال التدقيق في كميات البيانات الضخمة، باستخدام تقنيات تمييز النماذج (Pattern Recognition)، بالإضافة إلى التقنيات الرياضية والإحصائية، بهدف التنبؤ بالسلوك المستقبلي ووضع الحلول المناسبة للمشكلات قبل وقوعها في حال حدوثها أو من باب التنبؤ بها قبل وقوعها بهدف التطوير والتحديث بشكل عام في شتى المجالات [9].

في الواقع تُعتَبَر تقنية النظم الخبيرة (Expert Systems) من أهم فروع الذكاء الصنعي (Artificial Intelligence) وأكثرها تطوّراً، ويمكن تعريف النظام الخبير بأنّه عبارة عن برنامج حاسوبي يحتوي على خبرة الإنسان (Human Experience) ومَلكة التمييز (judgment) وقواعد التفكير (Rules of thinking)، والبديهية (Intuition)، وخبرات أُخرى لتقديم نصائح وحلول في تخصص أو مجال مُعين، ويمكن للنظام الخبير التخزين والمحافظة على الخبرة النادرة التي توجد عند عدد من الخبراء والتي يكون من الصعب استشارتهم في أيّ لحظة عند اللزوم[11]، وبعبارة أُخرى، إن النظام الخبير يحاول تقليد أو مُحاكاة الإنسان في تفكيره وطريقته في حل المشكلات في مجال معين [10].

يتميز النظام الخبير بقدرته الفائقة على تفاهمه مع المستخدم، وهذا التفاهم يتم عن طريق أسئلة محددة يوجهها النظام للمستخدم الذي يقوم بدوره بالإجابة عنها، ومن ثم يقوم النظام بتفسير المعلومات المنتقاة من تلك الإجابات وتحليلها ويصل إلى استنتاج قد يُساعد المستخدم على اتخاذ قرار أو القيام بإجراء يُسهم في حل المشكلة تحت الدراسة.

لقد تمَّ تطوير العديد من النظم الخبيرة في مجالات عدّة ولا يزال هناك العديد من المجالات التي تستخدم فيها النظم الخبيرة مما جعل هُناك تطوراً في أدوات تصميم وتطوير تلك النُظم لتلبية المتطلبات التقنية من حيث السرعة والكفاءة والقدرة على التوسع وغيرها وكذلك متطلبات المستخدمين من حيث طرق العرض والتعامل.

تهدف أشجار القرار إلى مساعدة متخذ القرار على التفكير المنظم في المشكلات المعقدة، ويركز على تحليل المشكلات الإدارية بطريقة عملية تسعى إلى تزويد متخذي القرار بالمعلومات المناسبة التى تساعده على التفهم والتبصُّر بالعوامل المؤثرة والظروف المحيطة بالقرار والنتائج المتوقعة للخيارات المتاحة [12].

تُعتبر نظم دعم اتخاذ القرار الذكية من تطبيقات الذكاء الصنعي [10]، فهي تساعد على اتخاذ القرار في كافة المجالات، الاقتصادية والزراعية وفي مجال الطقس وغيرها... [14].

في الواقع تُعتبَر أشجار القرار من أهم خوارزميات بناء نظم دعم اتخاذ القرار، والتي تعتبر من أهم أدوات تحليل القرار التي تساعد على اتخاذ القرار بالتوازي مع النماذج الاحتمالية وخوارزميات الاختيار [3].

يُمكن أن تُستخدم شجرة القرار لتمثيل القرارات وعمليات اتخاذ القرار بصرياً، وهذا يساعد في عملية تحليل القرار، و تَصِف أشجار القرار البيانات في عمليات التنقيب في البيانات (لكن ناتج شجرة التصنيف يُمكن أن يكون من المدخلات لعملية اتخاذ القرار).

تندرج خوارزميات شجرة القرار تحت فئة خوارزميات التعلم تحت الإشراف (Supervised Learning). والدافع العام لاستخدام شجرة القرار هو إنشاء نموذج

تدريب يمكن استخدامه للتنبؤ بفئة أو قيمة المتغيرات المُستهدفة من خلال تعلم قواعد القرار المُستَنتَجة من البيانات السابقة (بيانات التدريب) [8] .
باختصار فإن دور شجرة القرار هو مساعدتنا على اتخاذ قرارات جيدة أو قرارات في ظروف معينة [15].

في الواقع إنَّ العديد من الباحثين قد اهتموا بأشجار القرار والطرق التي تهتم ببنائها، وحاولوا تطوير خوارزمياتها المختلفة وتحسينها، فظهرت العديد من خوارزميات أشجار القرار التي سنتحدث عنها لاحقاً مثل خوارزمية ID3، و CART ، C4.5 ، C5.0 وغيرها [14].
ومن الجدير بالذكر أنَّ عملية إيجاد الشجرة الأفضل التي تُناسِب البيانات هي عملية NP (Nondeterministic Polynomial) والتي لا تُحَل بزمن نموذجي [15]، لذلك فإنَّ كل الطرق الموجودة هي طرق تقريبية نحاول اختيار طريقة تعطينا بناءً على التجريب على عدد كبير جداً من مجموعات البيانات (Data Sets) واعتماداً على دقة النتائج نستنتج أنها الطريقة الأمثل [15].

لنتكلم بدايةً عن الطرق التي تمَّ العمل عليها لبناء أشجار القرار، وهي طريقة تَعَلُّم خاضع للإشراف (Supervised Learning)، تُستخدم للتصنيف والانحدار (Regression). تُطبَّق عملية التصنيف عن طريق مجموعة من القواعد أو الشروط التي تحدد المسار الذي سيُتَّبَع ابتداءً من عقدة الجذر وانتهاءً بإحدى العقد النهائية التي تمثل الرمز للغرض المُصَنَّف، وينبغي عند كل العقد غير النهائية اتخاذ قرار حول مسار العقد التالية، وبسبب سهولة تطبيق أشجار القرار ووضوحها فهي منتشرة جداً ربما لأنها تشبه إلى حد ما طريقة تفكير الانسان لاتخاذ أي قرار أو التنبؤ به، لذا فإننا يجب أن نلاحظ عدة قواعد تتميز بها أشجار القرار الجيدة [14]:

1- يجب أن تتناسب شجرة القرار مع البيانات (Data)، بمعنى أن الصفات (Attributes) ذاتها يجب أن تعطي القرار ذاته.

2- من الأفضل أن يكون لشجرة القرار ارتفاعاً قليلاً، أي أنه لا يجب أن تعطي شجرة القرار عدداً من المسارات بقدر عدد البيانات (Data)، وإلا لن يكون هناك فائدة تُذكر لهذه الشجرة.

3- كلما كان ارتفاع الشجرة أقل وعدد عناصرها أقل كان ذلك أفضل، وكانت دقتها أعلى.

لذلك عند بناء شجرة القرار يجب ملاحظة ما يلي [9]:

أ- البيانات التي تصف حالة ما يجب أن تُحَلَّل إلى مجموعة من الصفات (Attributes) التي يُعَبَّر عنها في جدول حيث تملك كل صفة مجموعة من القيم (values) وقد تكون هذه القيم متقطِّعة أو مستمرة ويجب أن تكون قيم أي صفة ثابتة وليست متغيرة من حالة لأخرى (قيم صفة ما تارة متقطعة وأخرى مستمرة).

ب- يجب أن تكون الصفوف التي ستدل عليها العينات محددة مسبقاً، ويطلق على هذا المصطلح في مجال تعلم الآلة (Machine learning): التعلم بواسطة مُشرِف(supervised–learning). (حيث يمثِّل الصف القرار المتعلق بالصفات المرتبطة بعيّنة ما)

ج- يجب أن تكون الصفوف منفصلة (أو متقطعة) (discrete classes) (حيث تنتمي الحالة لصف معين).

د- يجب أن تكون البيانات كافية: حيث أن الاستقراء في صيغة شجرة القرار يعتمد على اختبارات إحصائية لذلك لابد من توافر عدد كافي من العينات حتى تكون هذه الاختبارات فعالة كما تتأثر كمية البيانات المطلوبة بعدة عوامل مثل عدد الصفات وعدد الصفوف وتعقيد نماذج التصنيف، فعندما تزداد هذه العوامل يجب أن تتوفر بيانات أكثر وذلك لبناء نموذج تصنيف أكثر وثوقية [11].

سنعرض في هذا البحث طريقة جديدة لبناء أشجار القرار تعتمد على دمج الخوارزمية الجينية (Genetic Algorithm) مع خوارزمية مستعمرة النمل (Ant Colony Algorithm).

كما سنقارن الطريقة المقترحة مع الطرق الأخرى وسنعرض النتائج التي توضح تحسين النتائج بشكل كبير عن الخوارزميات السابقة، مما يدل على أن هذه الخوارزمية يمكن أن تُعتَبَر أفضل من سابقاتها.

2. هدف البحث:

يهدف البحث إلى تطوير طريقة جديدة لبناء شجرة قرار فعالة وسريعة وأفضل من الطرق الموجودة حالياً، وذلك عن طريق استخدام منهجية هجينة تعتمد على دمج الخوارزمية الجينية بخوارزمية مستعمرة النمل، حيث نقوم بتمثيل الأشجار على شكل أفراد ضمن مجتمع ما لتطبيق الخوارزمية الجينية، ثم نجعل هؤلاء الأفراد يتأثرون بطريقة التعلُّم الخاصة بخوارزمية النمل، وبالتالي نستفيد من مميزات الخوارزميتين معاً، أي الاستفادة من ميزة وراثة الأجزاء الجيدة في الخوارزمية الجينية وميزة التواصل المجتمعي في خوارزميات النمل، وبالتالي يكون الهدف هو الوصول إلى شجرة القرار المثالية التي تُضاهي بقدراتها المنهجيات المتبعة سابقاً.

3. الدراسات السابقة:

- تم اقتراح المبدأ الاساسي للخوارزمية الجينية (Genetic Algorithm) عام 1975 من قبل العالم هولاند، وذلك وفقاً لنظرية دارون والتي تعتمد على مبدأ البقاء للأصلح. بحيث تعمل الخوارزمية اعتماداً على مبدأ أن الأفراد الذين هم أكثر عرضة للبقاء على قيد الحياة لفترة أطول هم الأشخاص الأفضل من حيث سماتهم الوراثية [14].

على مدى أجيال عديدة، تنتشر هذه المادة الوراثية الملائِمة لعدد متزايد من الأفراد، في الخوارزمية الجينية، يمثل كل فرد، بمعنى آخر كل "كروموسوم"

Chromosome ينتمي إلى المجتمع، حلاً محتملاً للمشكلة، وتمر الخوارزمية الجينية بعدة مراحل من أجل أداء وظيفتها، هذه المراحل هي: التهيئة والتقييم والاختيار والعبور (التزاوج) والطفرة [16].

- اخترع روس كوينلان J. Ross Quinlan خوارزمية الـ ID3 (Induction decision tree based on information Gain) لأول مرة عام 1975 والتي استخدمها لتوليد شجرة القرار من مجموعة من بيانات ثُنائية التفرع، وإنَّ جوهر فكرة خوارزمية ID3 هو العثور على طريقة تصنيف تقلل من درجة اضّطراب البيانات، ومن أجل هذا فإنها تستخدم فكرة الإنتروبي (إنتروبيا المعلومات Entropy). وكلما زادت قيمة الانتروبي زاد اضطراب البيانات وكلما انخفضت انخفض اضطراب البيانات، ولتصنيف البيانات نأمل بالتأكيد في الحصول على أقل انتروبي ممكن. بعد ذلك، وفي كل مرة نختار فيها صفة تصنيف يجب تقليل درجة اضطراب البيانات إلى أقصى حد، ويسمى مقدار هذا التخفيض "كسب المعلومات" أو معدل ربح المعلومات (Information Gain) [18].

بشكل عام نلاحظ أن خوارزمية ID3 تنحاز لإنتاج شجرة قرار صغيرة قدر الامكان، ووضع الصفات ذات الربح الأعلى قريبا من عقدة الجذر. لكنها لا تتعامل مع الصفات ذات القيم المستمرة أو الصفات التي تحوي على قيم مجهولة، كما أنها لا تَستخدِم تقنيات التقليم (Pruning) [18].

- وفي عام 1984 قدم بريمان Breiman خوارزمية CART (Classification And Regression Tree) لأول مرة. تَفترِض CART أن شجرة القرار عبارة عن شجرة ثنائية، وأن قيم ميزات العقدة الداخلية هي "نعم" و "لا"، والفرع الأيسر هو الفرع بالقيمة "نعم"، والفرع الأيمن هو الفرع بالقيمة "لا". تُعادِل شجرة القرار هذه التكعيب المتكرِّر لكل ميزة، وتقسيم مساحة الإدخال، أي مساحة الميزة، إلى وحدات محدودة، وتحديد توزيع احتمالية التنبؤ على هذه

الوحدات، أي الاحتمال الشرطي للمخرجات في ظل الظروف المعينة للإدخال [1].

تَستخدِم خوارزمية CART دليل جيني (Gini Index) (وهو مقياس إحصائي من مقاييس التشتت يهدف إلى تمثيل عدم المساواة بين قيم متغيرة)، لتقسيم عقدة إلى عقدتين فرعيتين، حيث تبدأ بمجموعة التدريب كعقدة جذر ثمَّ بعد تقسيم العقدة الجذرية بنجاح إلى قسمين تقوم بتقسيم المجموعات الفرعية باستخدام نفس المنطق وتقسيم المجموعات الفرعية مرة أُخرى ويتم تكرار هذا التقسيم حتى الوصول إلى عقد فرعية نقية أو أقصى عدد من الأوراق في شجرة نامية، ويطلق على هذه العملية اسم التَّقليم (Pruning) [1].

- في عام 1992 اقترح ماركو دوريغو فكرة خوارزمية مستعمرة النملACO (Ant Colony Algorithm) في اطروحته للدكتوراة [5]، وتنضم هذه الخوارزمية إلى عائلة خوارزميات ذكاء السرب[3].

  كانت الخوارزمية الأولى لدوريغو تهدف إلى البحث عن مسار أمثل في رسم بياني، استناداً إلى سلوك النمل الذي يسعى لإيجاد مسار بين المستعمرات ومصدر الغذاء. ثمَّ تنوَّعت الفكرة الأصلية منذ ذلك الحين لحل فئة أوسع من المشاكل العددية، ونتجة لذلك ظهرت عدة مسائل مستندة إلى جوانب مختلفة من سلوك النمل [7].

- وفي عام 1993 اقترح J. R. Quinlan تطويراً للخوارزمية السابقة وهي خوارزمية C4.5 التي تُعَدُّ امتداداً لخوارزميته الأولى ID3 وتُستَخدَم لإنشاء شجرة قرار هدفها التَّعَلُّم تحت إشراف، وتَعتمِد هذه الخوارزمية على طريقة Hunt's cls لبناء شجرة قرار من مجموعة عينات التدريب T (Training samples)، لكن بقيت مجموعة البيانات في عملية بناء الشجرة، في هذه الخوارزمية، تحتاج إلى مسح وفرز عدة مرات بالتسلسل مما قد يؤدي إلى عدم كفاءة الخوارزمية [15].

- ثمَّ لم يلبث Quinlan أن قام بتطويرها في عام 1997 إلى خوارزمية جديدة دعاها C5.0 لاعتمادها في البرمجيات التجارية [2]، ويمكنها الاستفادة من الحواسيب متعددة المعالجات (أي استخدام أمثل للموارد).

تَكُون شجرة القرار المُنشَأة بواسطة C5.0 سهلة في الفهم والنشر، وتَحُل هذه الخوارزمية العديد من انواع المشاكل الجيدة، كما يمكنها التَّعامل مع الخصائص العددية والاسمية بالإضافة إلى البيانات الناقصة، ويمكننا استخدامها على مجموعة كبيرة أو صغيرة من البيانات، إلا أنها في الكثير من الأحيان قد تؤدي بشجرة القرار إلى الانقسامات مع عدد كبير من مستويات الميزات، وقد يَصعُب في مرحلة من المراحل تجميع النموذج أو ملاحظة عدم ملاءمته، بالإضافة إلى أن التغيرات الصغيرة في بيانات التدريب قد تؤدي إلى تغيرات كبيرة في منطق القرار [2] .

- في عام 2004 استخدم الباحثان (Holden and Freitas) خوارزمية (Ant–Miner) في مجال التصنيف [6]، وأظهر البحث أن خوارزمية (Ant–Miner) كانت أكثر فعّالية من خوارزمية التصنيف (C5.0)، وأيضاً تمَّ التَّحقُّق من جدوى بعض التقنيات المعالِجة المبنية على اللغة لتقليل عدد السمات. أول تطبيق لخوارزمية أمثَلة (Optimizing) عناصر السرب في التصنيف كانت عام 2004 من قبل (Sousa et al) [13]، وقد اقترح هؤلاء الباحثون استخدام خوارزمية أمثَلة عناصر السرب بوصفها أداة جديدة للتنقيب في البيانات، ومثَّلوا ثلاثة نسخ لخوارزمية أمثلة عناصر السرب، وهم: خوارزمية أمثلة عناصر السرب المتقطعة (Discrete PSO)، وخوارزمية أمثلة عناصر السرب ذات الانحدار الخطي للأوزان (Linear Decreasing Weight PSO)، وخوارزمية أمثلة عناصر السرب المتقطعة (Constricted PSO)، حيث أنَّ PSO هو اختصار للجملة (Particle Swarm Optimization) [7].

وقورنت نتائجها مع الخوارزمية الجينية وخوارزمية شجرة الاستقراء.

- في عام 2009، حاول Hong Liu و Gengui Zhou تحسين أشجار القرار عن طريق دراسة مشكلة الحد الأدنى لامتداد الشجرة (MTS) (Minimum Tree Span) وتمَّ تطوير خوارزمية جينية ذاتية التكيف للتعامل مع هذه المشكلة. تَعتَمِد الطريقة التي اقتُرِحَت على استخدام ترميز للشجرة لتمكين مَعلَمات الإستراتيجية من التطور جنبًا إلى جنب مع العملية التطورية، لكن واجهت هذه الطريقة صعوبات في تحسين الشبكة التقليدية والتعامل معها [17].
- وفي عام 2020 تمَّ اقتراح طريقة لتحسين دقة التصنيف من قِبَل S. Rzheutskaya و A. Rzheutskiy و R. R. Salikhova وذلك بناءً على النموذج المفسَّر لأشجار القرار باستخدام خوارزمية جينية في مرحلة بناء الشجرة، وتمَّ تقديم وصف عام للخوارزمية وخصائص تنفيذ العمليات الجينية. لكن لم يتم مقارنة تنفيذ هذه الخوارزمية بالخوارزميات الأخرى ولم يتم اثبات كفاءتها بمقارنتها مع الطرق الأُخرى [16].

4. مشكلة البحث:

لم تُثبِت أيَّ خوارزمية من الخوارزميات الحالية لبناء أشجار القرار أنها خوارزمية أمثلية أو أنَّ أشجار القرار التي تُنتِجُها هي أشجار قرار فعّالة، لذا فإنَّ إيجاد شجرة قرار أمثلية هي مسألة NP (Nondeterministic Polynomial)، كما ذكرنا سابقاً [15].

لذلك نبحث عن حل يجعل الخوارزميات التي تَبني شجرة القرار المثالية فعالة، لأنَّ أشجار القرار ذات أهمية كبيرة ومنتشرة الاستخدام.

5. تعريف المسألة:

ليكن لدينا مجموعة من البيانات الجدولية، يُمَثِّل كل سطر عنصر من البيانات يسمى عينة (Sample). وتُمَثِّل الأعمدة –ماعدا العمود الأخير– الصِّفات (Attributes)، بحيث كل عمود يعبِّر عن صفة للعينات، يسمَّى العمود الأخير الهدف (Target)، ويُمَثِّل القرار المرتبط بكل عنصر عينة.

في مسألتنا ستكون كل القيم ثنائية (بوليانية) لإثبات صحة الخوارزمية، ولكن هذا لا يعني أن الخوارزمية لا تُطبَّق على بيانات متعددة القيم.

نريد بناء نموذج لشجرة القرار اعتماداً على هذه البيانات الموجودة بحيث تَعكِس هذه الشجرة ارتباط الهدف مع الخصائص بأحسن طريقة ممكنة، وتُحقِّق شروط الشجرة الأمثلية التي تحدثنا عنها سابقاً، وتكون قابلة للتعميم. أي أنَّ المسألة هي مسألة تعلُّم آلي (Machine Learning).

5. المنهجية المقترحة:

تَعتمِد المنهجية المقترحة على طريقة تَدمج بين خوارزمية النمل ACO والخوارزمية الجينية GA للاستفادة من خصائص كلا المنهجيتين.

إنَّ شجرة القرار تتألَّف من مجموعة من العقد، وكل عقدة تَرتبِط بإحدى الصفات. كل عقدة لها أب.

كما ذكرنا فإنَّ أشجار القرار تتميز عن بعضها بعدة معايير:

a) معايير قياس الأداء المرتبطة بالبيانات (حيث الأعلى هو الأفضل) (مثل: الدقة P (Precision)، الاستعادة R (Recall)، ومعيار f_1 (حيث f_1 هو المتوسط التوافقي للدقة P والاستعادة R، ويُحسب بالعلاقة:

$$f_1 = \frac{1}{\frac{\frac{1}{P}+\frac{1}{R}}{2}} = \frac{2PR}{P+R} \quad \text{.....} \quad (1)$$

b) ارتفاع الشجرة (حيث الأقل هو الأفضل).

c) عدد الأوراق (حيث الأقل هو الأفضل).

لذلك سنستخدم الجمع الموزون للقيم السابقة للحصول على جودة الشجرة. فإذا كان لدينا شجرة ارتفاعها الأقصى h وارتفاعها الحقيقي d وعدد أوراقها L. وعند حساب معامل f_1 (العلاقة (1)) كانت قيمته f. عندئذٍ تكون جودة هذه الشجرة G حيث:

$$G = a * f + b * \frac{h-d}{h} + c * \frac{2^{h-1} - L}{2^{h-1}} \quad \text{..... (2)}$$

حيث $a+b+c=1$ هي قيم اختيارية تعطي أفضلية لكل معيار (يمكن وضعها تجريبياً). ويفضل استخدام قيمة عالية لـ a لأن الأهمية هي للدقة.

سنقوم بتعريف دالة رياضية منطلقها هو الجداء الديكارتي لمجموعة الصفات، ومستقرها هو المجال $[0,1]$، ولندعوه $Ph(X,Y)$ تُعبِّر هذه الدالة عن الثقة بإمكانية أن تكون العقدة المرتبطة بالخاصية Y هي ابن للعقدة المرتبطة بالخاصية X. (ولنتذكر أننا أضفنا عقدة وهمية مرتبطة بخاصية وهمية للجذر). سيتم تحديد قيم هذه الدالة باستخدام خوارزمية النمل. نقوم بتوليد عدد من الأشجار بشكل عشوائي (الطريقة العشوائية كما سنذكر لاحقاً)، نقوم بتقييم كل شجرة من هذه الأشجار العشوائية بحسب الدالة G (العلاقة (2))، ونختار الشجرة الأفضل.

بما أن هذه الشجرة هي الشجرة الأفضل هذا يعني أن الارتباطات داخل الشجرة أدَّت إلى نتيجة جيدة، لذلك سنقوم بالاستفادة من هذه الشجرة لتعديل قيم الدالة $Ph(X,Y)$.

من أجل أي عقدتين أب وابن في الشجرة وليكونا مرتبطين بالصفتين X و Y ، نقوم بزيادة قيمة الدالة $Ph(X,Y)$ بمقدار ثابت تجريبي، ومن ثمَّ نقوم بعملية تطبيع (Normalization)، وبعد إعادة هذه العملية عدد كافٍ من المرات –آخذين بعين الاعتبار أنَّ التوليد العشوائي للأشجار مرتبط بالدالة $Ph(X,Y)$ كما سنذكر لاحقاً– فإنه ستتقارب قيم الدالة $Ph(X,Y)$ نحو قيم ثابتة تُحقِّق أفضل شجرة قرار ممكنة نوعاً ما.

لبناء الشجرة بشكل عشوائي متحيّز للدالة $Ph(X,Y)$ (أي بحيث يكون التوزُّع الاحتمالي مرتبط بالتابع $Ph(X,Y)$)، سنأخذ بعين الاعتبار الحقيقتين التاليتين:

1- كلما كانت الدالة $Ph(X,Y)$ ذات قيمة أعلى كان احتمال أن تكون العقدة Y هي ابن للعقدة X أكبر.

2- يجب الأخذ بعين الاعتبار في نظرية المعلومات (Information Theory) كلما كانت العقدة تحقق قيمة انتروبي (اعتلاج Entropy) أقل كلما كانت هذه العقدة أفضل ليتم اختيارها [4].

الآن سنُعَرِّف االدالة $Ph(X,Y)$ والتي هي احتمال أن تكون العقدة Y هي ابن للعقدة X. لإيجاد قيمة هذه الدالة سنقوم بما يلي:

لتكن T هي البيانات المرتبطة بالشجرة الفرعية المراد إيجاد جذر لها والناتجة عن تقسيم البيانات وفق العقدة X. تكون الدالة:

$$Ph(X,Y)=(Gini-index(T))*Ph(X,Y)$$

أصبح لدينا لكل عقدة احتمال بعد التطبيع. سنستخدم طريقة العشوائية المَوزونة لاختيار العقدة المناسبة. ونطبق هذا الكلام على كل عقد الشجرة حتى بناء الشجرة بشكل كامل.

لاستخدام الخوارزمية الجينية يجب أن نقوم بتمثيل كل شجرة بسلسلة من الأرقام، ويجب أن تكون هذه السلاسل ذات نفس الطول لكي تعمل الخوارزمية الجينية.

لتمثيل الشجرة على شكل سلسلة نقوم بما يلي:

نقوم بترقيم الصفات (Attributes) بدءاً من 0 . الرمز الأول ضمن السلسلة هو رقم الصفة المرتبطة بالجذر، من أجل كل شجرة فرعية نقوم بحذف الجذر وترقيم الصفات المتبقية من جديد وإعادة العملية.

يكون التمثيل النهائي هو دمج لتمثيل الجذر مع تمثيل الشجرة اليمنى ومن ثمَّ الشجرة اليسرى.

$$R(t)=\begin{cases} empty & t=empty \\ R(root)+R(left)+R(right) & otherwise \end{cases}$$

ولكن باعتبار أنَّ أشجار القرار تختلف بأعماقها وبُنيتها فإنَّ الطريقة السابقة لن تقوم بتوليد سلاسل ذات طول متساوي ولا يمكن أصلاً إعادة فك تشفيرها. لذلك سنقوم بما يلي:

سنقوم باعتبار كل أشجار القرار هي أشجار كاملة (Perfect)، أي أننا سنُضيف عقد غير ضرورية للأوراق لتتوسع الأشجار ويصبح عدد العقد متساوي في كل الأشجار. مع ملاحظة أنَّ هذه الإضافة لن تتسبَّب في تغيير بنية الشجرة الحقيقية، لأنَّ الوصول إلى ورقة في شجرة القرار عندما لايتم استخدام كامل الصفات هو مُكافِئ لاستخدام أي صفة من هذه الصفات بعد هذه الورقة وهو لا يضر إلا بزيادة ضئيلة جداً في التعقيد الزمني.

أصبح التمثيل لكل الأشجار متساوي الطول وأصبح التمثيل قابلاً للعكس، لأنَّ كل عقدة تُمَثَّل بدليل محدد، وهو ضِعف دليل الأب. (نظرية معروفة بتمثيل الأشجار الكاملة) [18].

بما أنَّ كل الأشجار أصبحت مُمثَّلة بشكل سلسلة من الأرقام وتُمَثِّل هذه السلاسل الكروموسومات في الخوارزمية الجينية، فيمكن تطبيق الخوارزمية الجينية التقليدية (من اختيار Selection وتزاوج Crossover وطفرة Mutation) ونُعِيدها حتى نصل للتقارب.

للاستفادة من الخوارزميتين معاً حيث تساعدنا خوارزمية النمل على التوليد العشوائي للأشجار بشكل يميل إلى الصحة، و تساعدنا الخوارزمية الجينية على تحسين الأشجار

الموجودة، سنقوم بتكرار توليد الأشجار وتزاوجها بشكل متتالي للاستفادة من خصائص الخوارزميتين، حيث نقوم بتوليد مجموعة من الأشجار بشكل عشوائي بطريقة خوارزمية النمل ثم نقوم بتحسين هذه الأشجار التي قمنا بتوليدها باستخدام الخوارزمية الجينية، وبعد عدد كافٍ من الدورات نقوم باختيار أفضل شجرة (حسب G) وتعديل قيم الدالة $Ph(X,Y)$. ونعيد العملية السابقة عدد كافٍ من المرات للوصول لأفضل شجرة.

فيما يلي نوجز خوارزميتنا المقترحة بالخوارزمية التالية:

Inputs: dataset: d; Rows: R; Attributes: A; Target: T;

number of generations: gn;

Outputs: decision tree;

Implementation

$Ph(X,Y)$ is Probability of node Y to be a child of node X

Initialize $Ph(X,Y)$ by using uniform distribution

Repeat until no enhancement

P is the current population with size n

Repeat n times

T = generate tree ($Ph(X,Y)$)

Add T to P

Repeat gn times

Pick tow individual T_1, T_2 using tournament selection

Encode T_1

Encode T_2

$(T_1^{'},T_2^{'})=crossover\ (T_1,T_2)$

Mutate $(T_1^{'})$

Mutate $(T_2^{'})$

add $(T_1^{'}, T_2^{'})$ to new population

P = new population

PT = find best tree

Update pheromone ($Ph(X,Y)$, PT)

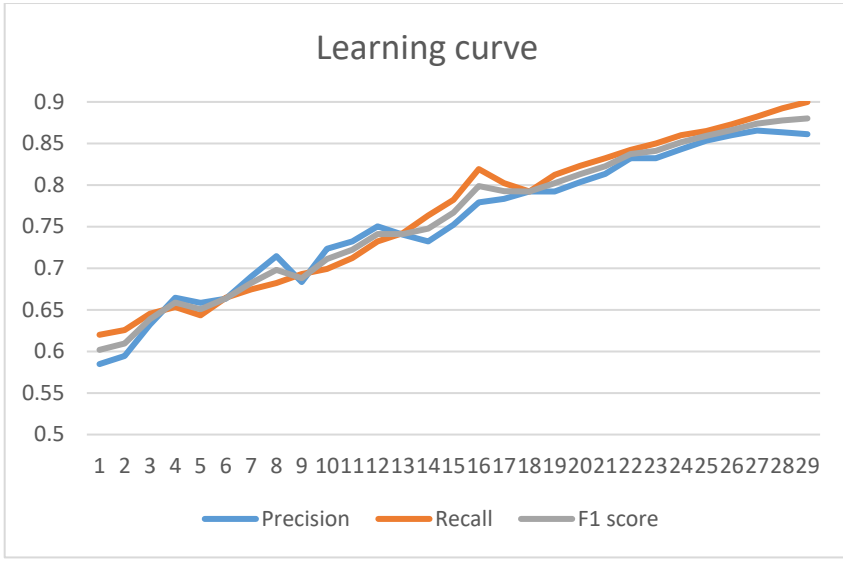
Print best tree

6. النتائج ومناقشتها:

قُمنا بتطبيق الخوارزمية السابقة باستخدام لغة بايثون وبرمجتها. لتجريب الخوارزمية السابقة حصلنا على مجموعة بيانات مرتبطة بأمراض القلب، الهدف منها تحديد ما إذا كان المريض المُحَدَّد بالصفات المطروحة مُرَشَّحاً للإصابة بقصور قلبي أم لا بحسب صفات طبية، مثل السن والضغط والتوتر الشرياني، ...

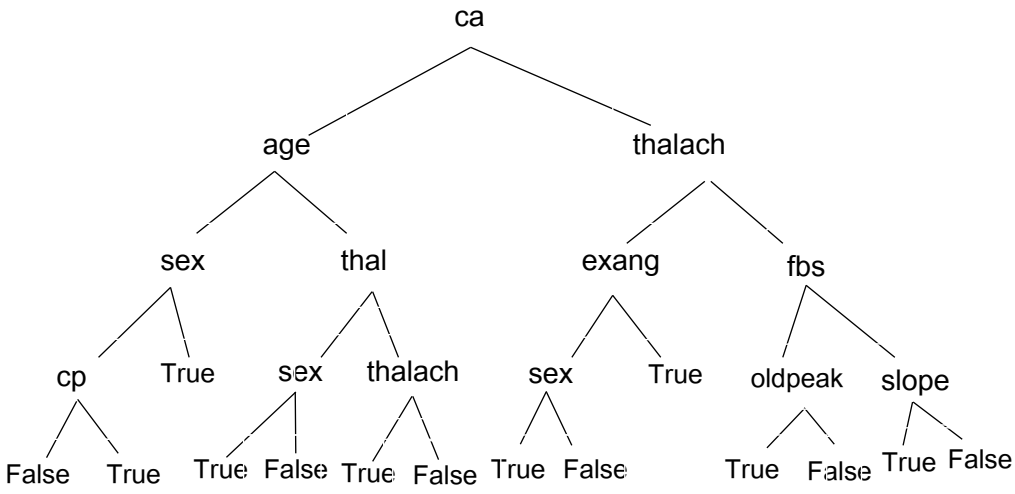
قمنا بتحويل كل البيانات إلى بيانات ثنائية (الضغط (مرتفع/ منخفض) – الجنس (ذكر/ أنثى)– العمر (كبير/ صغير) ...)، قمنا بتشغيل البرنامج وحصلنا على نتائج ممتازة.

يعرض الشكل (1) مخطَّط تَعَلُّم الخوارزمية ويُوَضِّح معايير القياس.



الشكل (1)

في الشكل (1)، كل خط يمثل إحدى الخوارزميات أثناء تنفيذ الخوارزمية خلال عدد الدورات. نلاحظ أنه على الرغم من تحسن الثقة في البداية في الخوارزمية الهجينة المقترحة إلا أنها لم تكن الأعلى، لكن بعد فترة أصبحت الثقة في خوارزميتنا هي الأعلى وتفوقت على الخوارزميات الأخرى.



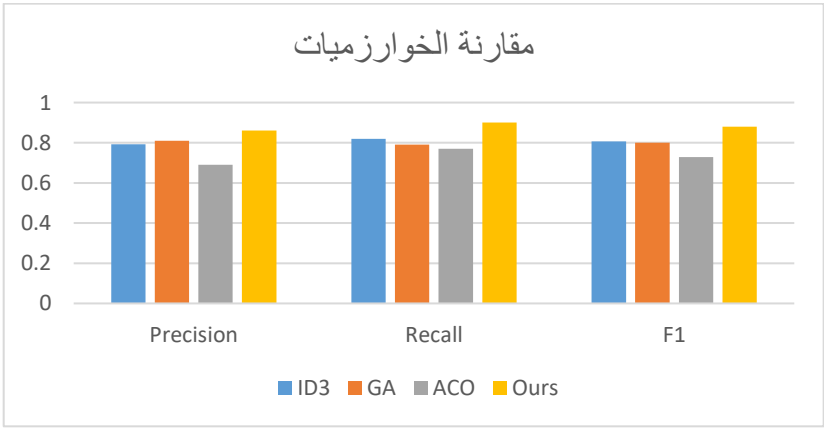
يعرض الجدول (1) النتائج المتعلِّقة بالشجرة المُستَقرَأة.

الجدول (1)

Method	Precision	Recall	F1 score	Leaf count	Height
GA+ACO	0.86122	0.8999	0.88013523	8	4
GA	0.81023	0.8092	0.809714672	9	4
ACO	0.69543	0.7923	0.74071127	10	4
ID3	0.79833	0.7822	0.790182693	8	4

تدل القيم المذكورة في الجدول (1) على فعَّالية الخوارزمية المقترحة.

لتوضيح كفاءة الخوارزمية المقترحة يجب مقارنتها مع المنهجيات السابقة، لذلك قمنا بتطبيق ID3 وخوارزمية النمل والخوارزمية الجينية كل على حدة، وتمت المقارنة معهم وفق معايير الأداء المقترحة. فوجدنا النتائج الموضحة بالشكل (2):



الشكل (2)

كما ذكرنا سابقاً، يعرض الشكل (2) نتائج تطبيق عدة خوارزميات على نفس مجموعة البيانات التي قمنا بالتجريب عليها، من حيث معايير قياس الأداء المُرتَبِطة بالبيانات، ويُظهِر تفوُّق خوارزميتنا بشكل ملحوظ وفق كل المعايير. نستنتج مما سبق أن الخوارزمية المُقتَرَحَة فعالة وأعطتنا نتائج أفضل.

7. الاستنتاجات والتوصيات:

عرضنا في هذه الورقة البحثية منهجية هجينة تمَّ فيها دمج خصائص خوارزمية النمل في توليد أشجار قرار واختيار الجيد منها عن طريق ميزة التواصل المجتمعي، وخصائص الخوارزمية الجينية في وراثة الأجزاء الجيدة. وأظهرنا النتائج التي حصلنا عليها وتمَّت مقارنتها مع الخوارزميات السابقة كل على حدة، وكما يبدو من النتائج أن الخوارزمية الهجينة المُقتَرَحَة جيدة وفعَّالة. ويجب اختبارها على نتائج أُخرى وتطويرها أكثر ربَّما بدمجها مع تقنيات أُخرى لخوارزميات أُخرى أيضاً.

8. المراجع:

1. Deepanker., 2019- Decision Tree with CART Algorithm. [Internet].
2. Czar Yobero., 2018- Determining Creditworthiness for Loan Applications Using C5.0 Decision Trees.
3. Dorigo M, Birattari M, Stutzle T., 2006- Ant colony optimization – IEEE Journals & Magazine [Internet].
4. Dorigo M., 2007- Ant colony optimization| Scholarpediam [Internet].
5. Gambardella L, Dorigo M., 2000- An Ant Colony System Hybridized with a New Local Search for the Sequntial Ordering Problem| INFORMS Journal on Computing [Internet].

6. Holden N. and Freitas A.A., 2004- Web Page Classification with an Ant Colony Algorithm, Conference: Parallel Problem Solving from Nature.

7. Holden N. P. and Freitas A. A., 2007- A Hybrid Pso/Aco Algorithm for Classification, GECCO Workshop on Particle Swarms.

8. Jurafsky D., Martin J. H., 2008- Speech and Language Processing, second edition, Prentice Hall.

9. Liu B., 2008- Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, Springer-Verlag , Berlin Heidelberg, New York.

10. Qi X. and Davison B.D., 2009- Web Page Classifcation: Features and Algorithms, ACM Computing Surveys.

11. Rao S.S., 2009- Engineering Optimization Theory and Practice.
12. Scime A., 2005- Web Mining Applications and Techniques.

13. Sousa T., Silva A., and Neves A.,2004- Particle Swarm based Data Mining Algorithms for classification tasks, Parallel Computing.

14. Z. Bandar, H. Al-Attar, K. Crockett, 1999- Genetic Algorithms For Decision Tree Induction.

15. I. S. Damanik, A. P. Windarto, 2019- Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm.

16. S. Rzheutskaya, A. Rzheutskiy, R. R. Salikhova, 2020- Applying a genetic algorithm to build a classification tree.

17. H. Liu, G. Zhou, 2009- Minimum Spanning Tree Problem Research Based on Genetic Algorithm.

18. G. Liang, 2005- A comparative study of three Decision Tree algorithms: ID3, Fuzzy ID3 and Probabilistic Fuzzy ID3.