تحسين أداء تصنيف تقانة آلة المتجه الداعم باستخدام الخوارزمية الجينية

أ.م.د. عمر صابر قاسم
قسم الرياضيات
كلية علوم الحاسوب والرياضيات
جامعة الموصل

محمد علي محمد الوزان
قسم الرياضيات
كلية علوم الحاسوب والرياضيات
جامعة الموصل

تاريخ استلام البحث : ٢٠١٨\٤\٢٥
تاريخ قبول البحث : ٢٠١٨\٨\٢٧

ABSTRACT

In this research, the genetic algorithm was proposed as a method to find the parameters of support vector machine, specifically the σ and c parameters for kernel and the hyperplane respectively. Based on the Least squares method, the fitness function was built in the genetic algorithm to find the optimal values of the parameters in the proposed method. The proposed method showed better and more efficient results than the classical method of support vector machine which adopts the default or random values of parameters σ and c in the classification of leukemia data .

Key words: Genetic Algorithm, Support Vector Machine, Parameter Selection.

الملخص

اقترح في هذا البحث الخوارزمية الجينية(Genetic algorithm) طريقة لإيجاد المعلمات Parameters)) لتقانة آلة المتجه الداعم (Support vector machine) وتحديدا المعلمات σ و c اللتان تمثلان على التوالي معلمة النواة (Kernel)، ومعلمة المستوى الفاصل Hyperplane))، وبالاستناد إلى طريقة المربعات الصغرى ((Least squares بنيت دالة اللياقةFitness function)) في الخوارزمية الجينية التي يحدد بها أفضل قيم للمعلمات في الطريقة المقترحة، وبالمقارنة مع الطريقة الاعتيادية لتقانة آلة المتجه الداعم التي تعتمد قيم افتراضية أو عشوائية للمعلمات σ و c، تبين أن الطريقة المقترحة ذات نتائج أفضل واكفأ من الطريقة الاعتيادية في تصنيف بيانات مرض اللوكيميا.

الكلمات المفتاحية: الخوارزمية الجينية، آلة المتجه الداعم، اختيار المعلمات.

1. مقدمة

تعد تقانات الذكاء الاصطناعي Techniques artificial intelligence من أهم الطرائق المستخدمة في العديد من المجالات (التطبيقات) وفي مقدمتها المجال الطبي. إذ تمتلك مرونة في التعامل مع هذه التطبيقات المختلفة وتحديد الحالات المرضية وتصنيفها، كما تتميز هذه التقانات بسهولة كبيرة في التعامل مع البيانات وتحليلها وتشخيصها بعدة إجراءات رياضية تعتمد على خوارزميات مبنية وفق نماذج رياضية متعددة. إذ إن العديد من التطبيقات وخصوصاً التطبيقات الطبية التي تعتمد التشخيص الجزيئي Diagnosis molecular للجينات تحتوي كميات كبيرة من البيانات تصل إلى الآف القراءات اللازمة لكل حالة مرضية مما يؤدي إلى صعوبة كبيرة في التعرف عليها من دون الأدوات والخوارزميات الحاسوبية التي تقوم بهذه المهمة.

إن العديد من التقانات الذكائية مثل الخوارزمية الجينية Genetic algorithm وآلة المتجه الداعم Support vector machine والمنطق المضبب Fuzzy logic وغيرها تعتمد على معلمات Parameters أساسية تدخل في تكوين هذه التقانات وتؤثر على نحو مباشر في عملية التصنيف، فمثلاً تقانة آلة المتجه الداعم يدخل في تكوينها معلمات أساسية مثل c التي تمثل معلمة المستوى الفاصل Hyperplane و σ التي تمثل معلمة النواة Kernel . إذ تعطى قيم اختيارية وعشوائية لهذه المعلمات أثناء بناء تقانة آلة المتجه الداعم وتركيبها، إلا أن اختيار هذه المعلمات بصورة عشوائية قد يؤدي إلى أخطاء في عملية التصنيف في كثير من الحالات إن لم تختر على نحو صحيح.

إن عملية تصنيف البيانات في التطبيقات الطبية التي تعتمد على بيانات كبيرة مثل مرض اللوكيميا من النوعين سرطان الدم الليمفاوي الحاد Acute lymphocytic leukemia (ALL) وسرطان الدم النخاعي الحاد Acute myeloid leukemia (AML) تحتاج إلى دقة في اختيار التقانة وإلى أمثليه في اختيار معلمات هذه التقانة كونها تؤثر في عملية التصنيف على نحو مباشر نتيجة التعقيد الحاصل في طبيعة البيانات المستخدمة. وقد اقترح في هذه الدراسة آلية تعتمد الخوارزمية الجينية في اختيار أفضل قيم لمعلمات تقانة آلة المتجه الداعم σ و c لتستخدم في بناء تركيب التقانة والحصول على امثل قيم المعلمات بالنسبة للبيانات المستخدمة [1].

2. الدراسات السابقة

يركز في هذه الفقرة على أهم تفاصيل الأعمال البحثية المختلفة المتعلقة بهذه الدراسة بخصوص تقانتي الخوارزمية الجينية GA، وآلة المتجه الداعم SVM. إذ قام الباحثان Vladimir Cherkassky, Yunqian Ma في عام ٢٠٠٣ بعملية اختيار المعلمات الزائدة في آلة المتجه الداعم SVM واختيرت المعلمة التحليلية مباشرة من بيانات التدريب بدلا عن أساليب إعادة اخذ العينات المستخدمة على نحو شائع تطبيقات SVM. وفي عام ٢٠١٣ توصل الباحثان Ilha Ilhan, Gulay Tezel باستخدام الخوارزمية الجينية إلى التنبؤ بقيم المتغيرات مع تقانة آلة المتجه الداعم وطرحها باسم GA-SVM لتحديد علامات SNP (تعدد أشكال النوكليوتيدات الأحادية) التي تتضمن ملايين المتغيرات في الجينوم البشري. إذ تم تحسين قيمة C معلمة آلة المتجه الداعم بوساطة الخوارزمية الجينية، وتظهر النتائج التي تم الحصول عليها أن هذه الطريقة يمكن أن توفر أفضل دقة في تحديد SNP مقارنة بالطرائق الأخرى.

3. الخوارزمية الجينية

إن ظهور الخوارزمية الجينية (GA) بدأ رسمياً في العام ١٩٧٥ على يد العالم جون هولاند John Holland في جامعة Michigan، وسميت هذه الخوارزمية بالجينية لاعتمادها على مبدأ عمل الكروموسومات والجينات الوراثية في الكائنات الحية للتوصل إلى أفضل الحلول. إذ تعد من أهم الأساليب الحديثة في مجال التقانات الذكائية، إذ ظهرت أهمية استخدامها في حل مسائل معقدة فضلاً عن حل المسائل الصعبة في مختلف العلوم، كما استخدمت منذ بداية تطورها مع موضوعات مهمة مثل الشبكات العصبية Neural network والروبوتات Robotics والامثلية Optimization، وكذلك حل مسائل التشفير وكسر الشفرة وغيرها، إذ تهتم الخوارزمية الجينية عموماً بكيفية إنتاج أفراد جديدة تمتلك صفات معينة (مرغوبة أو غير مرغوبة)، بالتداخل أو التعديل أو التبديل الذي يحصل على المجموعات الموروثة بهدف تكوين هذه الأفراد، إذ ركز عليها الباحثون كثيراً،

وذلك لسهولتها ولكونها لا تحتاج إلى معرفة عميقة بالخصائص والتفاصيل الرياضية مثل قابلية الاشتقاق والتقعر التي لا يمكن توافرها أحياناً في بعض التطبيقات [2،3].

1.3. دالة اللياقة

تعد دالة اللياقة أهم المكونات الأساسية التي تعتمدها الخوارزمية الجينية، إذ تحسب فيها لياقة كل كروموسوم (فرد) في المجتمع، كما تسمى أحياناً بدالة الهدف للخوارزمية الجينية التي بها تحدد الأفراد التي يتم اعتمادها في حل المسائل. هناك العديد من الآليات التي تعتمدها الخوارزمية الجينية في دالة اللياقة، منها مسائل تعتمد على تعظيم Maximization المسألة (أي الحصول على أعلى قيمة لدالة اللياقة) وتستخدم في التطبيقات التي يتم فيها حساب الإنتاج والربح، أما المسائل التي تعتمد على التصغير Minimization فإن الهدف منها هو إيجاد الحل ذي القيمة الأصغر لدالة الهدف وتستخدم في مسائل حساب الخطأ والجذور التي تحقق المعادلات، إذ يقيم كل كروموسوم أو فرد عن طريق دالة رياضية مخصصة لإعطاء قيمة تعكس لياقة هذه الكروموسومات لتختار كأفضل الحلول [4].

2.3 العمليات الجينية

تعتمد آلية الخوارزمية الجينية على ثلاثة مراحل مهمة للمعالجة وهي [5، 6]:

١. الاختيار Selection.
٢. التزاوج أو التقاطع Crossover.
٣. الطفرة Mutation.

1.2.3 الاختيار Selection

هي عملية اختيار الزوج المناسب لكل كروموسوم من المجتمع الابتدائي، إذ يختار الأزواج لأجل التزاوج وإنتاج جيل جديد، إذ إن عملية الاختيار تحدد كيفية اختيار الأفراد الذين سيبقون في المرحلة المقبلة ومن الأساليب المتبعة لاختيار الأزواج هي:

١. انتقاء عجلة الروليت Roulette.
٢. اختيار المجموعات Tournament.
٣. الانتقاء النسبي Proportional selection.

2.2.3 التزاوج أو التقاطع Crossover

هي عملية إنتاج كروموسومات جديدة ذات صفات أفضل من صفات الأبوين ومن أهم أنواع طرائق التزاوج (التقاطع) في الخوارزمية الجينية:

١. تقاطع (تزاوج) ذو النقطة الواحدة Single point crossover.
٢. التقاطع (التزاوج) ذو النقطتين Two point crossover.
٣. التقاطع المنتظم Uniform crossover.

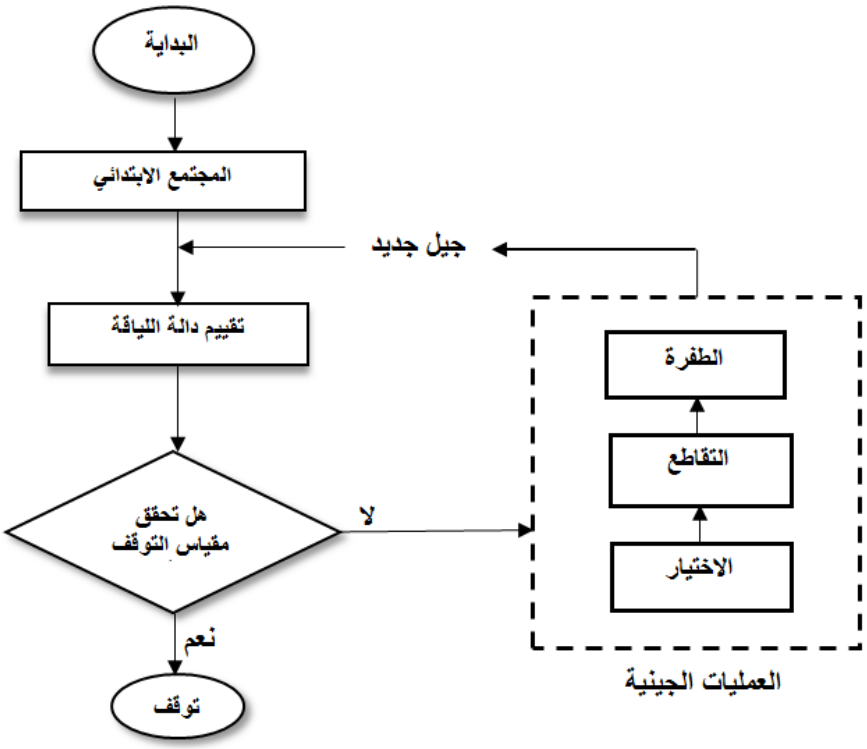
3.2.3 الطفرة Mutation

الطفرة هي عبارة عن إجراء بسيط لتغيير أو إبدال قيمة معينة ضمن الكروموسوم الناتج من عملية الاختيار، وإن القيمة المحددة المختارة لغرض إبدالها تختار عشوائياً، والطفرة عادة هي إجراء يجري على الفرد لغرض تحسين صفاته الجينية في المجتمع وهنالك عدة أنواع للطفرة ومنها:

١. تغيير الرتب Bit inversion.
٢. الإضافة والطرح Adding or subtract.
٣. عكس الجين Order changing.

3.3 مقياس التوقف

هنالك عدة مقاييس تعتمد في الخوارزمية الجينية لتحقيق شرط التوقف في حالة عدم الوصول للحل المثالي أو الأمثل، منها تنفيذ الخوارزمية عدد الأجيال المطلوبة ويحددها الباحثون، وكذلك تحديد وقت مستغرق لتنفيذ الخوارزمية أو الحصول على أحسن قيمة لدالة اللياقة في المجتمع الذي تم الوصول إليه عندما تكون أقل من قيمة اللياقة التي حددت في حالة (Minimize) أو تتوقف عندما تكون أحسن قيمة لدالة اللياقة أكبر من قيمة اللياقة التي حددت في حالة (Maximize) ويفحص مقياس التوقف للخوارزمية الجينية بعد تكوين كل جيل جديد لنرى إذا تم الوصول إلى الحل المرغوب [4، 5] الشكل (1) يبين آلية عمل خطوات الخوارزمية الجينية.



الشكل (1) يمثل مخططاً انسيابياً يوضح خطوات عمل الخوارزمية الجينية

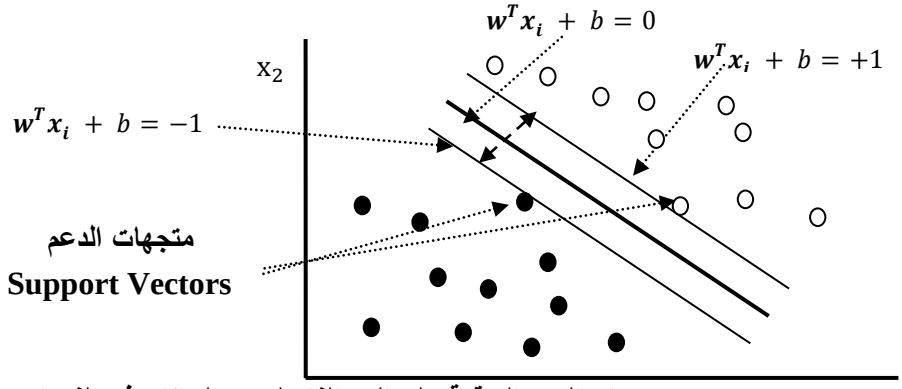
4. آلة المتجه الداعم

إن تقنات آلة المتجه الداعم SVM تعد من أهم التقانات المستخدمة في تصنيف البيانات، إذ تعتمد على عوامل متعددة ومتغيرات تؤثر على نحو مباشر أو غير مباشر على إيجاد الحل النهائي. فمثلاً أنموذج تقانة آلة المتجه الداعم يعتمد على عدد من المعلمات الأساسية مثل المستوى الفاصل Hyperplane، ومضاريب لاكرانج Lagrange multipliers، التي تؤثر على نحو كبير على دقة عملية التصنيف، إذ إن البيانات الأساسية في فضاء الإدخال (Input Space)، تصنف على وفق الأنموذج الرياضي الآتي [4، 7]:

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \geq +1 \qquad for \quad d_i = +1\ , \qquad i = 1,2,\dots N \qquad (1)$$

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1 \qquad for \quad d_i = -1, \qquad i = 1,2,\dots N \qquad (2)$$

$\boldsymbol{w}$: متجه الأوزان Weights ، $\boldsymbol{x}$ تمثل متجه الإدخال Input Vector، b تمثل قيمة التحيز Bias، و d تمثل قيمة الإخراج. كما يمكن ملاحظة معادلات الحدود في المستوى من الشكل (2) الآتي:



(2) : يوضح متجهات الدعم الواقعة على الحد الفاصل بين البيانات في الفضاء

من الشكل (2) نلاحظ أن معادلة المستوى الفاصل Hyper plane تكتب بالشكل الآتي:

$$\boldsymbol{w}^T\boldsymbol{x}_i + b = 0 \qquad (3)$$

إن البيانات القريبة أو التي تقع على حدود الحد الفاصل تسمى متجهات الدعم أو المساندة Support vectors. كما يمكن حساب المسافة بين النقاط في المستوى ومعادلة المستوى الفاصل من خلال العلاقة الآتية:

$$d(\boldsymbol{w}, b, \boldsymbol{x}_i) = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{i}} + \mathrm{b}|}{\|\mathbf{w}\|} \qquad (4)$$

وبعد عدد من الإجراءات والتحويلات الرياضية، يتم إيجاد قيم كل من متجه الأوزان المثالي (w^*) والتحيز المثالي (b^*) ومن بعدها تحسب دالة التصنيف الآتية:

$$f(x) = I(\boldsymbol{w}^* \cdot \boldsymbol{x} + b^*) \qquad (5)$$

إذ إن $\boldsymbol{w}^*$: تمثل الوزن المثالي ، b^* : تمثل قيمة التحيز المثالية و sgn: تمثل القرار النهائي لانتماء $(\boldsymbol{x})$ لأحد الأصناف Classes [8, 9].

1.4. عملية الفصل غير الخطي في الة المتجه الداعم

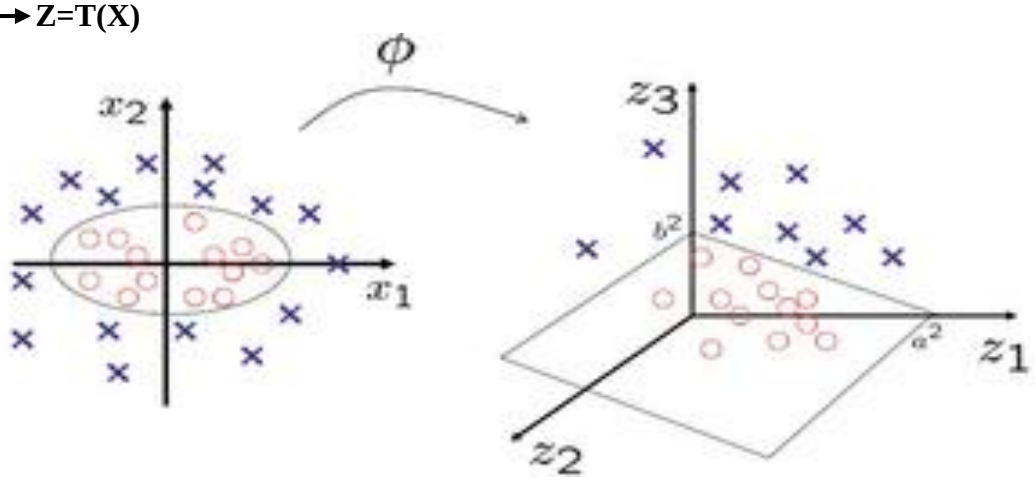
إن عملية التصنيف في أغلب التطبيقات تعتمد أسلوب الفصل غير الخطي Nonlinear لأن البيانات التي صنفت تكون عموماً غير خطية، لذلك يعتمد على هذه الآلية في معظم تطبيقات تقانة آلة المتجه الداعم التي تأخذ

إجراءات رياضية في تكوين دالة الفصل النهائية، ومن أهم هذه الإجراءات دالة النواة Kernel trick التي تقوم بتحويل البيانات من فضاء منخفض البعد Low dimensional إلى فضاء عالٍ البعد High dimensional من أجل تصنيفها [10، 11].

1.1.4. دالة النواة

في بعض الحالات لا يمكن فصل البيانات بوساطة الحد الفاصل Hyperplane، لذلك يفضل استخدام تقانة Kernel trick، التي تقوم بتحويل بيانات المدخلات من الخطية إلى بيانات ذات مساحة ابعاد عالية (بيانات غير خطية)، ويكون التحويل الجديد قابلاً للفصل الخطي بوساطة الحد الفاصل Hyperplane، الشكل (3) الآتي يبين كيفية تحويل البيانات من الخطية إلى غير الخطية باستخدام Kernel Trick [10]:

T:X⟶ Z=T(X)



الشكل (3): يوضح عملية تحويل البيانات من الخطية إلى غير خطية باستخدام تقانة Kernel Trick

$$T: X = (x_1, x_2, \dots, x_n) \rightarrow Z = \left(z_1 = T(x_1), z_2 = T(x_2), \dots, z_n = T(x_n)\right).$$

2.1.4. أنواع دوال النواة

أهم أنواع دوال النواة التي تستخدم في مسائل التصنيف [12، 13].

نوع الدالة	الصيغة الرياضية
Gaussian (RBF) kernel	$k(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$
Polynomial kernel	$k(x_i, x_j) = (1 + x_i^t x_j)^d$
Linear kernel	$k(x_i, x_j) = x_i^t x_j$
Monomial	$k(x, y) = [x^t y]^m$

3.1.4. خصائص دالة النواة

هناك العديد من الخصائص التي تعتمدها دالة النواة منها:

١. **تكون متناظرة Symmetric:**

$$k(x,z) = \big(\emptyset(x)\,.\,\emptyset(z)\big) = \big(\emptyset(z).\emptyset(x)\big) = k(z,x)$$

٢. **تحقق شرط Cauchy-Schwarz Inequality:**

$$k(x,z)^2 = (\emptyset(x).\emptyset(z))^2 \le \|\emptyset(x)\|^2\|\emptyset(z)\|^2$$
$$= \big(\emptyset(x).\emptyset(x)\big)\big(\emptyset(z).\emptyset(z)\big)$$
$$= k(x,x)k(z,z)$$

5. أسلوب الـ Cross-Valind

يعد أسلوب الـ Cross –Valind من أهم الأساليب المستخدمة في عملية تصنيف البيانات كونه يعطي آلية محايدة وذات كفاءة عالية في بناء المصنفات المختلفة، إذ يعتمد هذا الأسلوب على تقسيم البيانات إلى مجموعتين أساسيتين وهما مجموعة التدريب Training set ومجموعة الاختبار Testing set، اذ تمثل مجموعة التدريب مجموعة البيانات التي سوف تستخدم لغرض تكوين الأنموذج بينما مجموعة الاختبار تمثل مجموعة البيانات التي سوف تستخدم لغرض التنبؤ بالأنموذج الذي تم تكوينه من مجموعة بيانات التدريب، إذ يعتمد أسلوب ال Cross-Valind على تقسيم البيانات إلى k من المجموعات الجزئية وأغلب الأحيان تكون متساوية أو قريبة من التساوي وأن كل مجموعة سوف تحتوي على بيانات للمتغير المعتمد وعلى بيانات المتغير المستقل وهو ما يسمى بأسلوب ال k-fold Cross-Validation، إذ يتم تكوين الأنموذج من (k-1) من المجموعات والمجموعة المتبقية سوف تستخدم مجموعة اختبار للتنبؤ بهذا الأنموذج وعادة ما تكون قيمة k ضمن الفترة ($3 \le k \le 10$)، وعلى هذا الأساس يعاد في كل مرة تكوين أنموذج بإدخال مجموعة التدريب السابقة إلى (k-1) وتسحب مجموعة جديدة أخرى للتنبؤ وهكذا إلى k من المرات [14، 15].

6. الطريقة المقترحة GA_SVM

اقترح في هذا البحث طريقة مهجنة بين الخوارزمية الجينية GA وآلة المتجه الداعم SVM لإيجاد أفضل المعلمات (c التي تمثل معلمة المستوى الفاصل Hyperplane و σ التي تمثل معلمة النواة Kernel) لتقانة آلة المتجه الداعم SVM، إذ تبنى دالة اللياقة في الخوارزمية الجينية بالاعتماد على طريقة المربعات الصغرى Least square، وبعد إجراء عدة خطوات وبتكرارات متعددة تقوم الخوارزمية الجينية باختيار أفضل قيم للمعلمات c و σ ووضعها في تركيب تقانة آلة المتجه الداعم وبنائها ومقارنتها مع الطريقة الاعتيادية التي تستخدم في الغالب قيم عشوائية للمعلمات c و σ في تركيب التقانة.

لقد قورنت الطريقة المقترحة GA_SVM التي تعتمد على المعلمات σ و c التي تم الحصول عليها من الخوارزمية الجينية مع تقانة آلة المتجه الداعم الاعتيادية SVM وذلك من خلال التطبيق على بيانات مرض اللوكيميا، واستخدم أسلوب الـ Cross –Valind في توزيع مجاميع الاختبار والتدريب وبمعدل (50) تكراراً، يوزع في كل تكرار البيانات على نحو مختلف بين مجاميع التدريب والاختبار ليتم تأكيد مدى فعالية الخوارزمية المقترحة GA_SVM في الحصول على المعلمات c و σ وتبين من خلال حساب متوسط مربع الخطأ MSE أن

الخوارزمية المقترحة ساهمت على نحو فعال في تحسين دقة التصنيف لبيانات اللوكيميا مقارنة بالخوارزمية الاعتيادية.

1.6. الخوارزمية المقترحة لتصنيف أمراض اللوكيميا ALL وAML تكون على النحو الآتي:

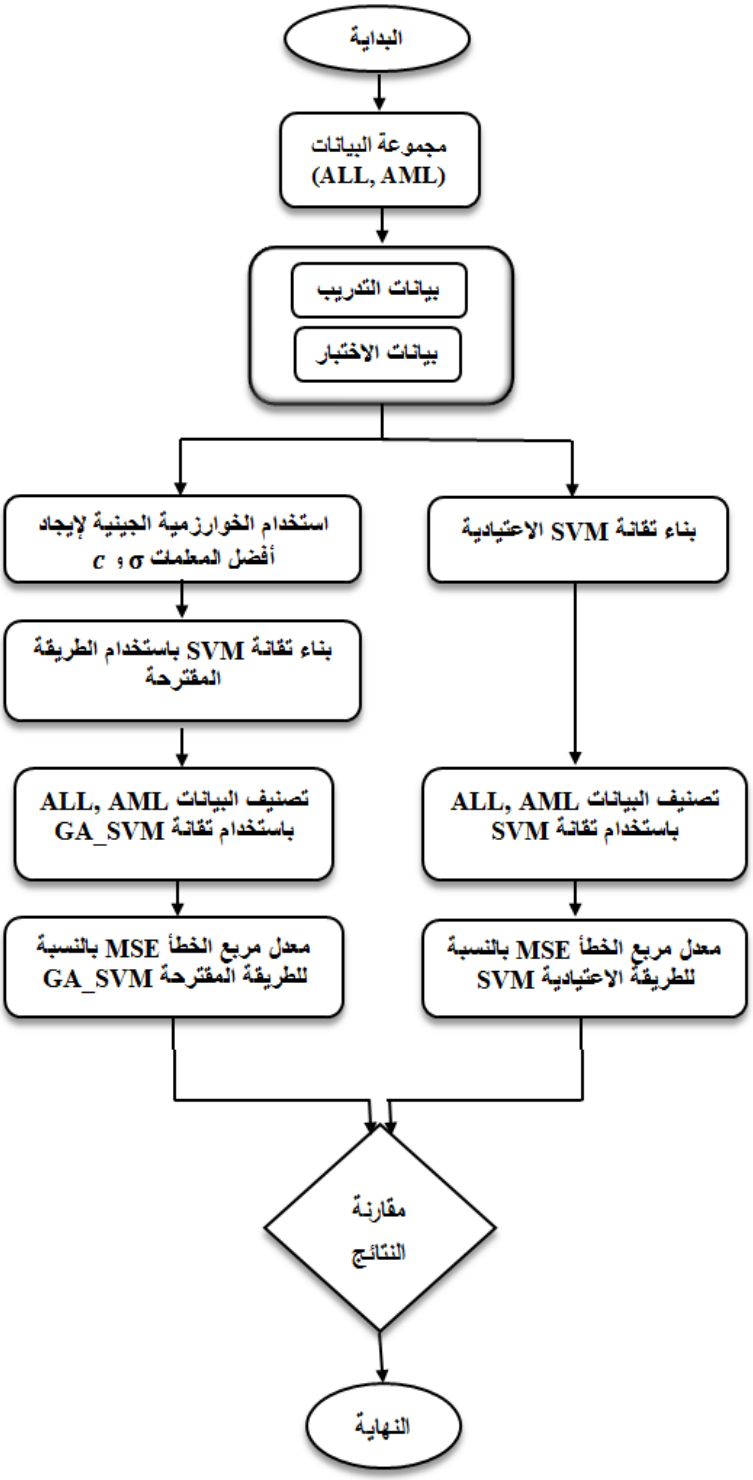
١- تهيئة بيانات الإدخال بشكل مصفوفة A صفوفها تمثل عدد الحالات المرضية الكلية وأعمدتها تمثل عدد الجينات التي يقاس تأثيرها والتي تمثل الميزات بالنسبة لبيانات اللوكيميا ALL وAML.

٢- استخدام أسلوب ال Cross Valind لتقسيم البيانات المرضية إلى مجموعات تدريب Training وأخرى اختبار Testing.

٣- استخدام الخوارزمية الجينية GA ومن خلال مفهوم دالة اللياقة في الحصول على أفضل المعلمات σ و c.

٤- استخدام أفضل المعلمات σ و c التي تم الحصول عليها من الخوارزمية الجينية GA في تركيب تقانة آلة المتجه الداعم SVM وبنائها لأجل تصنيف بيانات التدريب والاختبار.

مقارنة نتائج الطريقة المقترحة GA_SVM التي بنيت بالاعتماد على أفضل المعلمات σ و c مع تقانة آلة المتجه الداعم SVM الاعتيادية.



الشكل (4) مخطط يوضح آلية عمل الطريقة المقترحة

7. النتائج التجريبية

طبقت الخوارزمية المقترحة GA_SVM على بيانات اللوكيميا والمأخوذة من موقع البيانات العالمي (UCI Machine Learning Repository)، إذ إن عدد الحالات للبيانات المرضية هي ٧٦ حالة والتي تصنف إلى نوعين (ALL, AML) وكل حالة تحتوي على ٧١٢٩ ميزة (Feature) أو جين لكل حالة مرضية، يعد كل منها قياس مدى فعالية جين معين وتمثل بأعداد حقيقية، تم الحصول على هذه الميزات عن طريق المصفوفة الدقيقة

Microarray. إذ استخدم أسلوب الـ Cross –Valind لتقسيم البيانات بصورة مختلفة (أي عندما تكون قيمة الـ K-fold=5 وعندما تكون قيمة الـ K-fold=10 وعندما تكون قيمة الـ K-fold=15) وذلك للتحقق من النتائج على نحو دقيق، إذ قورنت الطريقة المقترحة مع الطريقة الاعتيادية على النحو الآتي:

الجدول (1): يبين مقارنة بين تقانة SVM الاعتيادية والخوارزمية المقترحة GA_SVM عندما (k-fold=5)

التقانة المستخدمة	متوسط مربع الخطأ بالنسبة لبيانات التدريب	متوسط مربع الخطأ بالنسبة لبيانات الاختبار
الطريقة المقترحة GA_SVM	0	1.8333e-02
الطريقة الاعتيادية SVM	0	3.3833e-02

الجدول (2): يبين مقارنة بين تقانة SVM الاعتيادية والخوارزمية المقترحة GA_SVM عندما (k-fold=10)

التقانة المستخدمة	متوسط مربع الخطأ بالنسبة لبيانات التدريب	متوسط مربع الخطأ بالنسبة لبيانات الاختبار
الطريقة المقترحة GA_SVM	0	5.0000e-03
الطريقة الاعتيادية SVM	0	1.5159e-02

الجدول (3): يبين مقارنة بين تقانة SVM الاعتيادية والخوارزمية المقترحة GA_SVM عندما (k-fold=15)

التقانة المستخدمة	متوسط مربع الخطأ بالنسبة لبيانات التدريب	متوسط مربع الخطأ بالنسبة لبيانات الاختبار
الطريقة المقترحة GA_SVM	0	0
الطريقة الاعتيادية SVM	4.0000e-03	8.0000e-03

اتضح من النتائج في الجداول (1) و (2) و (3) أعلاه أن الخوارزمية المقترحة GA_SVM تعطي نتائج أفضل من الخوارزمية الاعتيادية SVM في جميع الحالات التي اعتمدت كبيانات اختبار، في حين أن بيانات التدريب

كانت مطابقة في الجداول الثلاثة للخوارزمية المقترحة، لكن الخوارزمية الاعتيادية اخطأت في بعض التصنيفات في الجدول (3)، مما يؤكد كفاءة الخوارزمية المقترحة وذلك بمقياس متوسط مربع الخطأ MSE.

8. الاستنتاجات والتوصيات Conclusions and Recommendation

تضمنت هذه الدراسة اقتراح الخوارزمية الجينية GA كطريقة لإيجاد افضل قيم للمعلمات σ و c في تقانة آلة المتجه الداعم SVM، إذ طبقت الخوارزمية المقترحة GA_SVM لتصنيف بيانات اللوكيميا من النوعين ALL, AML ومن خلال التطبيق العملي على البيانات المرضية تبين بأن نسبة التصنيف للخوارزمية المقترحة GA_SVMتفوق نسبة التصنيف لتقانة SVM القياسية وذلك من خلال مؤشر مقياس متوسط مربع الخطأ MSE، والمبينة خلال الجداول (1) و (2) و (3)، مما يدل على كفاءة الخوارزمية المقترحة GA_SVM مقارنة بتقانة آلة المتجه الداعم الاعتيادية.

كما نوصي بدراسة تطوير تقانة آلة المتجه الداعم SVM بإيجاد أفضل المعلمات لها باستخدام التقانات الذكائية الأخرى مثل خوارزمية السرب Particle swarm optimization والمنطق الضبابي Fuzzy logic والشبكات العصبية Neural networks، كما يمكن اعتماد الخوارزمية المقترحة GA_SVM في تصنيف العديد من التطبيقات المختلفة التي تحتوي بيانات معقدة وكبيرة.

المصادر

[1] Li, X.Z. and J.M. Kong, Application of GA–SVM method with parameter optimization for landslide development prediction. Natural Hazards and Earth System Science, 2014. 14(3): p. 525-533.

[2] Kozeny, V., Genetic algorithms for credit scoring: Alternative fitness function performance comparison. Expert Systems with Applications, 2015. 42(6): p. 2998-3004.

[3] Oreski, S. and G. Oreski, Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications, 2014. 41(4): p. 2052-2064.

[4] Min, S.-H., J. Lee, and I. Han, Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Systems with Applications, 2006. 31(3): p. 652-660.

[5] Motieghader, H., et al., A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Informatics in Medicine Unlocked, 2017. 9: p. 246-254.

[6] Sastry, K., D.E. Goldberg, and G. Kendall, Genetic Algorithms. 2014: p. 93-117.

[7] Steinwart, I. and A. Christmann, Support vector machines. 2008: Springer Science & Business Media.

[8] Chen, F.-L. and F.-C. Li, Combination of feature selection approaches with SVM in credit scoring. Expert Systems with Applications, 2010. 37(7): p. 4902-4909.

[9] Zhou, L., K.K. Lai, and L. Yu, Credit scoring using support vector machines with direct search for parameters selection. Soft Computing, 2008. 13(2): p. 149-155.

[10] Bellotti, T. and J. Crook, Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 2009. 36(2): p. 3302-3308.

[11] Danenas, P. and G. Garsva, Selection of Support Vector Machines based classifiers for credit risk domain. Expert Systems with Applications, 2015. 42(6): p. 3194-3204.

[12] Huang, C.-L., M.-C. Chen, and C.-J. Wang, Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 2007. 33(4): p. 847-856.

[13] Shin, K.-S., T.S. Lee, and H.-j. Kim, An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 2005. 28(1): p. 127-135.

[14] Maris, F., et al. Support vector Machines-Kernel algorithms for the estimation of the water supply in cyprus. in International Conference on Artificial Neural Networks. 2010. Springer.

[15] Shao, X. and M.-a. Sun, Predicting Gene Expression Noise from Gene Expression Variations, in Transcriptome Data Analysis. 2018, Springer. p. 183-198.