مجلة جامعة طرطوس للبحوث والدراسات العلمية_ سلسلة العلوم الأساسية المجلد (6) العدد (5) 2022

Tartous University Journal for Research and Scientific Studies –Basic Sciences Series Vol. (6) No. (5) 2022

تحسين خوارزمية شجرة القرار بالاستفادة من النقاط الشاذة

محمّد مصطفى حجّوز*

(تاريخ الإيداع 5/6/ 2022 – تاريخ النشر18/10/ 2022)

□ ملخّص □

تعدُّ أشجار القرار من خوارزميات التصنيف الأكثر استخدامًا في تقنية التنقيب في البيانات، فالمصنفات الناتجة عنها دقيقة وسهلة الاستخدام والفهم. وكما هو الحال في معظم نماذج التصنيف الأخرى التي تتجاهل الاستثناءات في شجرة القرار إذ تعدُّ خارج مجال التصنيف (بيانات غير مفيدة). لكن في الحقيقة يمكن الاستفادة من هذه الاستثناءات لاكتشاف جزء هام من المعرفة وقد تكون ثمينة ومطلوبة لتعديل قراراتنا في ظروف نادرة غير عادية.

ونظرًا لأن الاستثناءات تتعلق بعدد ضئيل من الحالات، فمن الصعب ايجادها بسهولة لأن خوارزمية التعلم تركّز على اكتشاف المعرفة دون الشاذ منها. من هنا قام هذا البحث ايجاد خوارزمية محسّنة للتعلم في شجرة القرار تأخذ بعين الاعتبار الاستثناءات الناتجة عنها. وتم تجربة عمل الخوارزمية المقترحة من خلال مجموعة من البيانات التي أُختيرت بعناية، وطُبقت على مجموعتين من البيانات التي حُصل عليها من مستودع التعلم الآلي. وفي نهاية البحث وجدنا أن الخوارزمية تكتشف عدة استثناءات من خلال مجموعات البيانات التجريبية، ويمكن من خلالها اتخاذ الإجراءات المناسبة.

الكلمات المفتاحية: التصنيف، الاستثناء، شجرة القرار المحسّنة.

*مدّرس محاضر –في جامعة البعث– بكلية العلوم الثانية –سورية M.Hajjouz@gmail.com

مجلة جامعة طرطوس للبحوث والدراسات العلمية_ سلسلة العلوم الأساسية المجلد (6) العدد (5) 2022

Tartous University Journal for Research and Scientific Studies –Basic Sciences Series Vol. (6) No. (5) 2022

Enhancing the decision tree algorithm by taking advantage of anomalies

D.muhammad Mustafa hajour*

(Received 5/6/2022.Accepted 18/10/2022)

□ABSTRACT □

Decision trees are one of the most widely used classification algorithms in data mining, the classifiers generated by them are accurate and easy to use and understand. As in most other classification models that ignore exceptions in the decision tree because they are considered outside the scope of the classification (not useful data). But in fact, we can take advantage of these exceptions to discover an important piece of knowledge that may be valuable and needed to modify our decisions in unusually rare circumstances.

Since exceptions are related to a small number of cases, it is difficult to find them easily because the learning algorithm focuses on discovering knowledge rather than the anomaly. Hence, in this research, an improved algorithm for learning in the decision tree was found that takes into account the resulting exceptions. The work of the proposed algorithm was tested through a set of carefully selected data, and it was applied to two sets of data obtained from a machine learning repository. At the end of the research.

We found that the algorithm detects several exceptions from experimental datasets, from which appropriate decisions can be take.

Keywords: Classification, Exception, Enhanced Decision Tre**e.**

*lecturer at al-baath university-faculty of science-syria

١ – المقدمة:

تعدُّ تقنية التنقيب في البيانات من التقنيات الجديدة نسبياً، التي يمكن البحث فيها لإيجاد تقنيات تصنيف فعّالة من أجل اكتشاف المعرفة من قواعد البيانات [١]. تستخدم خوارزمية شجرة القرار على نطاق واسع لاحتوائها على أدوات قويّة لتصنيف البيانات، حإذ تضم فيها سلسلة من الاختبارات البسيطة المرتبطة ببعضها منطقيا، يقوم كل اختبار بتقييم الواصفة Attribute مقابل قيمة حدية معينة أو واصفة راجحة مقابل مجموعة من معلماتها.

تعدُّ مصنّفات شجرة القرار أكثر فائدة من النماذج الأخرى كالشبكات العصبية وآلات متجه الدعم، فيما يتعلق بقابلية التفسير من شكل If–Then فالقواعد المشتقة من أشجار القرار أسهل بكثير لصانعي القرار من أوزان نموذج الشبكة العصبية أو المعادلات الرياضية المملة المستخدمة لإيجاد الحد الأقصى في خوارزمية آلة المتجهات الداعمة.

تُبنى شجرة القرار من نماذج بيانات التدريب، فهي تتشكل من خلال عملية تكرارية مطوّلة من الأعلى إلى الأسفل، يبدأ فيها تكوين الشجرة مع عقدة الجذر التي تحتوي على بيانات التدريب بالكامل. وتعدُّ الواصفة التي تقسّم فيها بيانات التدريب الواصفة الأفضل في عقدة الجذر. ومن ثم تُقسّم مجموعة بيانات التدريب إلى بيانات مختلفة ومجموعات فرعية تفي بمعايير تقسيم واصفة التقسيم هذه. بعدها تكرّر هذا الإجراء لكل مجموعة فرعية بشكل تدريجي حتى تقع جميع الأمثلة في مجموعة فرعية وفي فئة واحدة.

في مرحلةٍ لاحقة تُنفّذ إصدارات مختلفة من مصنفات أشجار القرار مثل CART وID3 وID4 وID5 وC4.5 وC5.0 [10]. وتعتمد هذه الاصدارات على التطور التدريجي للمعالجة، إلا أنه لا يمكن لأي من هذه الاصدارات إيجاد الحالات الاستثنائية الموجودة في البيانات.

تعتبر الاستثناءات النادرة في القاعدة العامة مجال عمل مفيد جداً، على سبيل المثال، في قاعدة البيانات المتعلّقة بالطيور، يمكن للمصنف استخلاص قاعدة مفادها أن كل طائر يمكنه الطيران، علماً أن هناك أنواع من الطيور لا يمكنها الطيران مثل الكيوي والنعام والبطريق. فإذا تجاهلنا مثل هذه الأنواع من الحالات الاستثنائية سنخسر الكثير من المعرفة. من هنا قمنا في بحثنا هذا بتحسين خوارزمية شجرة القرار التي تقوم على تحديد مجموعات الاستثناءات وتصنيفها في الفئة المناسبة، تحدّد هذه الاستثناءات أثناء بناء شجرة القرار من بيانات التدريب. استخدم لإنجاز البحث عينة من مجموعة البيانات وثلاث مجموعات من البيانات الحقيقية لإظهار كيفيّة اكتشاف قواعد التصنيف الكاملة والأكثر دقة.

نُظّمت ورقة البحث على النحو الآتي:

- المقدمة.
- الدراسة المرجعية، وبينّا فيها المواضيع ذات الصلة بالعمل في هذا المجال.
- تحديد عينة مجموعة البيانات التي تستخدم لاكتشاف الاستثناءات والمصممة خصيصاً لهذا الهدف.
- النتائج التجريبية والتحقق من صحتها في بعض مجموعات البيانات الحقيقية.
- الخاتمة والمراجع.

٢ – الدراسة المرجعية:

من المنطقي أن يرتاح صانعوا القرار مع النماذج التي تكون سهلة الفهم. وتعدُّ أشجار القرار مفهومة أكثر من مثيلاتها الأخرى كالشبكات العصبية وآلات متجه الدعم. علاوة على ذلك فإن خوارزميات شجرة القرار تحظى بشعبية واسعة منذ نشائها الأولى كخوارزميات SPRINT, C4.5, CART, SLIQ مقارنة مع خوارزميات التعلم الأخرى[٥].

يظهر هذه البحث الإصدارات اللاحقة من أشجار القرار مثل C4.5 و C5.0 والتي لديها معدل خطأ أقل وسرعة أكبر. إضافة إلى ذلك، هناك العديد من الطرق مثل خوارزمية نمو الأشجار السريعة Fast Tree– Growing Algorithm [٩] وخوارزمية تقسيم البيانات Data Partitioning [١٠] وخوارزمية التوازي Parallelization [٢] كل هذه الخوارزميات أُقترحت من قبل الباحث كريك Gehrke في مطلع العام ٢٠٠٠ لتطوير خوارزميات سريعة وقابلة لتطوير بناء أشجار القرار [٨]. كما اقترحت العديد من الاستراتيجيات لتحسين خوارزمية أشجار القرار[٦] وكان التركيز منصبّاً على التعديلات الصغيرة لنماذج التصنيف المتاحة. وقد تم الاهتمام أيضاً بالعديد من المصنفات المستندة إلى المجموعة Ensemble–Based Classifiers [٧] والتي تؤدي إلى تحسّن طفيف بخصوص الدقة. حقيقة لا يوجد أبحاث كافية لتحسين عملية التعلم باستخدام أشجار القرار من أجل اكتشاف المعرفة فيما يخص النقاط الشاذة، وقد تم التركيز فقط على تحسين الدقة دون التطرق إلى الاستثناءات والتي قد تكون مهمة ومفيدة في اتخاذ القرار.

كما طُوّرت العديد من تقنيات التنقيب في البيانات لاكتشاف الأشياء المثيرة للاهتمام من عدد كبير من القواعد المكتشفة [٩]. والتي تستند إما على بعض الإجراءات المثيرة للاهتمام لتصفية القواعد غير المهمة أو بالاعتماد عليها بناءً على معرفة مجال المستخدم لتحديد القواعد غير المتوقعة [٤]. كما اقترح كل من كومبتون Compton وجيسن Jansen قواعد متتالية لاكتشاف المعرفة من خلال قواعد النموذج أسفلاً من أجل تحديد الاستثناءات وفق العديد من المستويات. أما سوزوكي Suzuki وزملاؤه فقاموا بتصنيف الاستثناءات في فئات مختلفة [٩]، وأهتموا بالتنقيب في استثناءات النموذج من خلال قاعدة الأزواج والثلاثيات لنمذجة التبعية في عمل التنقيب في البيانات التي يمكن أن تتخذ العديد من السمات مثل تسميات الصف، وقد تمكنوا من إيجاد عدة استثناءات لكنها تبدو غير مناسبة للتحليل والاستنتاج. كما اكتشف نوع آخر من الاستثناءات تسمى بقواعد الإنتاج الخاضعة للرقابة Censored Production Rules (CPRs) (CPRs) باستخدام النهج التطوري من قبل Vashishtha وBala وآخرون في عام ٢٠١٣ [١]. وفي عام ٢٠١٦ تم ايجاد خوارزميات جينية Genetic Algorithms لاكتشاف قواعد التصنيف وقواعد التصنيف الضبابية مع استثناءات داخل الصف وبين فئات صفوف مجموعات البيانات التي تحتوي على سمات اسمية Nominal ومستمرة Continues [3]. أُقترح نهج الخوارزمية الجينية لاكتشاف قواعد التصنيف الخاضعة للرقابة الضبابية، لم تستخدم أي من الخوارزميات السابقة أشجار القرار كطريقة تعلم في اكتشاف الاستثناءات، لكن استخدم نوع مختلف لمعالجة الاستثناءات التي تحسّن خوارزميات استقراء القاعدة لحل الروابط التي تظهر في حالات معينة ضمن البيانات أثناء عملية إنشاء القاعدة[١١]، بينما يركز بحثنا على اكتشاف الاستثناءات التي تتوقف بموجبها الحالات العامة ونسير في طريق الاستثناء إلى النهاية.

٣ – إيجاد الاستثناءات:

لا تهتم خوارزمية شجرة القرار التقليدية عند بنائها في إيجاد الاستثناءات بل تتجاهلها وتعدّها حالات شاذة[١٢]. لهذا قمنا في هذه البحث بتحسين خوارزمية شجرة القرار من أجل التركيز على ايجاد هذه الاستثناءات كجزء أساس من قواعد التصنيف الناتجة عن نموذج شجرة القرار. ولتبسيط الأمور وجعلها قابلة للفهم، سوف تستخدم عينة من مجموعة بيانات Dataset الفطر (مجموعة واصفات صغيرة من مجموعة بيانات الفطر العديدة). تحتوي مجموعة بيانات الفطر على ٧ سمات و٣١ حالة مع صنفين للفطر: "صالح للأكل" Eatable و"سام" Poisonous. وبافتراض أن واصفة "الرائحة" Odor هي الواصفة الراجحة في تمييز نوع الفطر، لذاا يبدأ التقسيم عند عقدة الجذر على هذه الواصفة.

يوضح الشّكل (١) الجداول الفرعية (من A إلى F) لمجموعة البيانات النموذجية. يُظهر كل جدول فرعي جزءاً من بيانات العينة التي تعطيها أزواج الواصفة والقيمة المميزة على طول الفروع المختلفة لشجرة القرار مع قيم "الرائحة" وهي "a"، "c"، "f"، "l"، "n"، "p"، يشير الرقم الوارد بين قوسين أسفل كل جدول فرعي إلى عدد المثيلات في الجدول الفرعي ذو الصلة. جميع الجداول الفرعية من "A" إلى "F" هي أقسام صحيحة من البيانات فإما أن تنتمي إلى فئة "صالحة للأكل" أو إلى فئة "سامة" باستثناء الجدول الفرعي "D"، الذي يحتوي على عشرة حالات من أزواج قيمة الواصفة "Odor=n" منها ٨ حالات تنتمي إلى فئة يكون فيها الفطر "صالح للأكل"، وحالتين تنتمي إلى فئة يكون فيها الفطر "سام". قد تكون آخر حالتين حالات شاذة، أو بعبارة أخرى قد تكون استثناءات قيمّة ومفيدة.

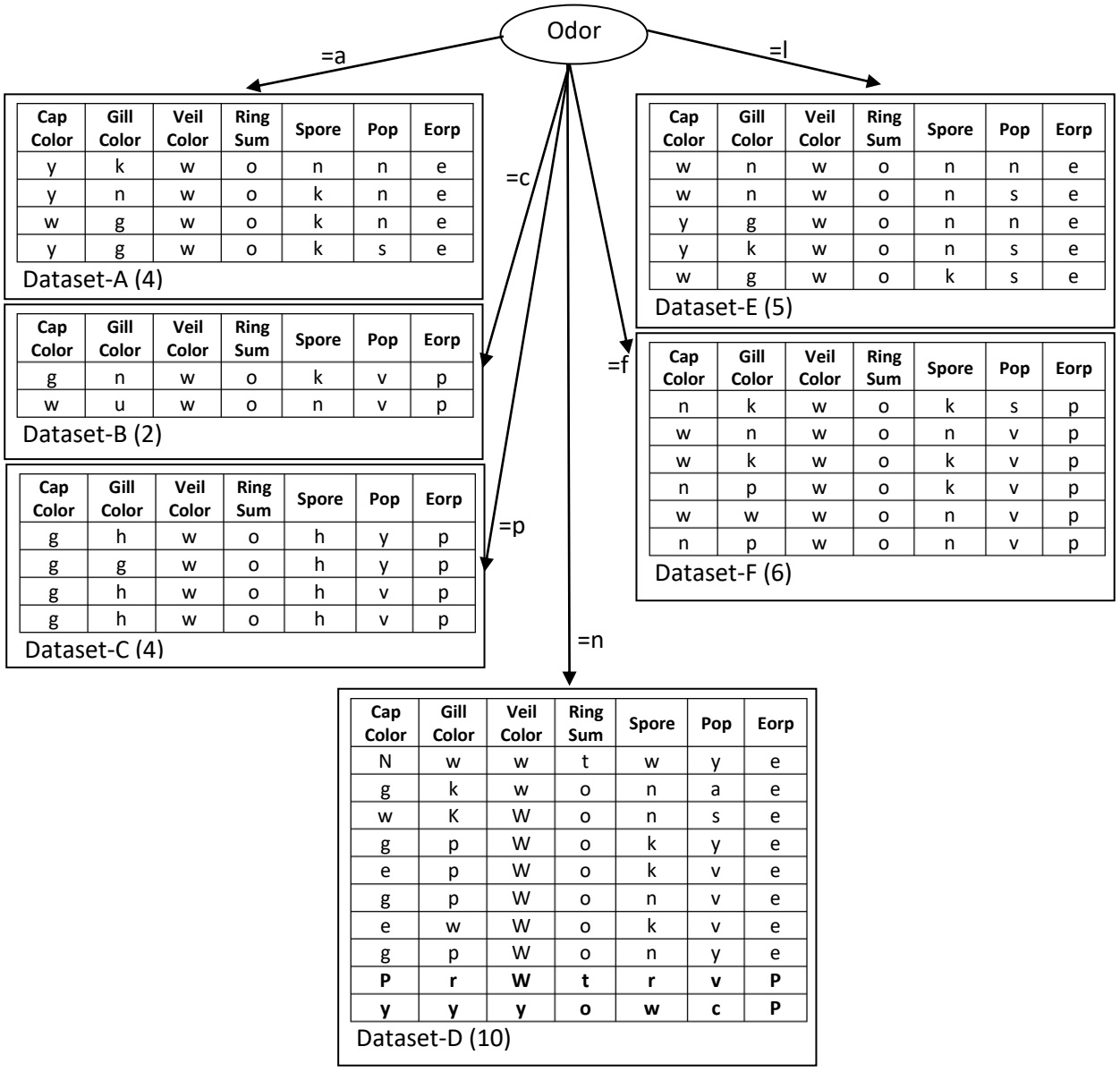
تتوقف عملية بناء الشجرة المعتادة عند هذه العقدة وتنتهي بقاعدة "إذا كانت (Odor=n) فيكون الفطر عندها "صالح للأكل" If the odor=n Then Mushroom = "eatable" ذلك لأن المزيد من التقسيم قد ينتج عنه شجرة قرار قد لا تناسب بيانات التدريب وتكون قوة تعميمها منخفضة. من الواضح أن نموذج شجرة القرار هذا لا بد أن يخطئ في تصنيف بعض الحالات التي تحتوي على أزواج قيمة واصفة "Odor = n" من بيانات الاختبار.

يوضح الجدول (١) بعض واصفات الفطر التي يمكن من خلالها تمييز الفطر السام من غيره واختصاراتها.

الجدول (١) بعض واصفات الفطر التي نحتاجها في بحثنا (Attribute Information)

1	cap-color لون الغطاء	brown=n	buff=b	cinnamon=c	gray=g	green=r	
		بني	برتقالي	قرفي	رمادي	أخضر	
		pink=p	purple=u	red=e	white=w	yellow=y	
		وردي	أرجواني	أبيض	أحمر	أصفر	
2	odor الرائحة	almond=a	anise=l	creosote=c	fishy=y	foul=f	
		اللوز	اليانسون	الكريوزوت	مريب	سيئة	
		musty=m	none=n	pungent=p	spicy=s		
		متعفن	لا شيء	لاذع	حار		
3	gill-color لون الخيشومية	black=k	brown=n	buff=b	chocolate=h	gray=g	green=r
		أسود	بني	برتقالي فاتح	شوكولاتة	رمادي	أخضر
		orange=o	pink=p	purple=u	red=e	white=w	yellow=y
		برتقالي	وردي	أرجواني	أحمر	ابيض	اصفر
4	veil-color لون الحجاب	brown=n	orange=o	white=w	yellow=y		
		بني	برتقالي	أبيض	أصفر		
5	ring-number عدد الحلقات	none=n	one=o	two=t			
		لا شيء	واحد	اثنان			
6	Spore color لون البوغ	black=k	brown=n	buff=b	chocolate=h	green=r	
		أسود	بني	برتقالي فاتح	شوكولاتة	أخضر	
		orange=o	purple=u	white=w	yellow=y		
		برتقالي	أرجواني	أبيض	أصفر		

6	Population(pop) التجمعات الفطرية	abundant=a	clustered=c	numerous=n	scattered=s	several=v	solitary=y
		وفيرة	مجمعة	عددية	متناثرة	متعددة	انفرادية
٧	Eatable صالح للأكل ام سام	Eatable=e	Poisonous=p				
		صالح للأكل	سام				



Cap Color	Gill Color	Veil Color	Ring Sum	Spore	Pop	Eorp
y	k	w	o	n	n	e
y	n	w	o	k	n	e
w	g	w	o	k	n	e
y	g	w	o	k	s	e

Cap Color	Gill Color	Veil Color	Ring Sum	Spore	Pop	Eorp
w	n	w	o	n	n	e
w	n	w	o	n	s	e
y	g	w	o	n	n	e
y	k	w	o	n	s	e
w	g	w	o	k	s	e

Cap Color	Gill Color	Veil Color	Ring Sum	Spore	Pop	Eorp
g	n	w	o	k	v	p
w	u	w	o	n	v	p

Cap Color	Gill Color	Veil Color	Ring Sum	Spore	Pop	Eorp
n	k	w	o	k	s	p
w	n	w	o	n	v	p
w	k	w	o	k	v	p
n	p	w	o	k	v	p
w	w	w	o	n	v	p
n	p	w	o	n	v	p

Cap Color	Gill Color	Veil Color	Ring Sum	Spore	Pop	Eorp
g	h	w	o	h	y	p
g	g	w	o	h	y	p
g	h	w	o	h	v	p
g	h	w	o	h	v	p

Cap Color	Gill Color	Veil Color	Ring Sum	Spore	Pop	Eorp
N	w	w	t	w	y	e
g	k	w	o	n	a	e
w	K	W	o	n	s	e
g	p	W	o	k	y	e
e	p	W	o	k	v	e
g	p	W	o	n	v	e
e	w	W	o	k	v	e
g	p	W	o	n	y	e
P	**r**	**W**	**t**	**r**	**v**	**P**
y	**y**	**y**	**o**	**w**	**c**	**P**

الشّكل (١) شجرة القرار لمجموعة بيانات الفطر

يُمثّل نموذج شجرة القرار إما كمجموعة من القواعد أو على شكل شجرة، لكن بالرغم من وجود حسابات غير صحيحة لهذا التصنيف، إلا أنه لا يمكن القيام بأي عمل لمنع حدوث مثل هذه المشكلة، كما أن هناك حاجة إلى إجراء جديد لدمج اكتشاف الاستثناءات الصحيحة في عملية بناء شجرة القرار. تبين الخوارزمية

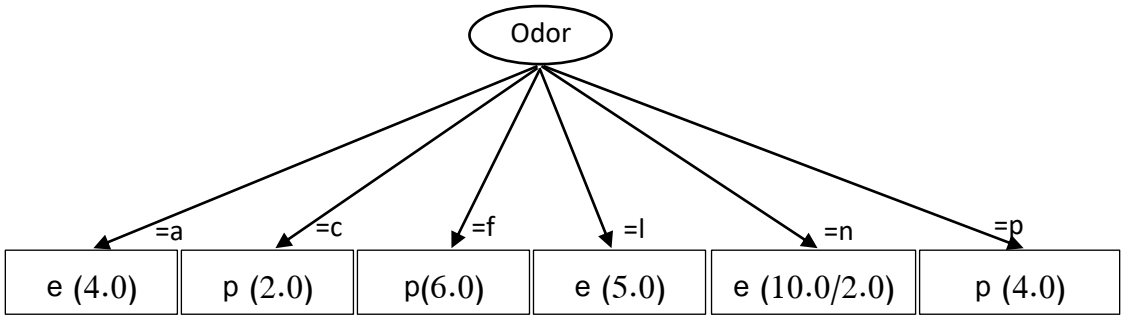
المحسّنة الآتية في الشكّل (٢) مخطط شجرة القرار المعدّل الذي يكتشف الاستثناءات. إذ تقّسم الخوارزمية المحسّنة مجموعة البيانات المتشابهة إلى فئات مختلفة وتكتشف الشروط الاستثنائية أيضاً.

خوارزمية إنشاء شجرة القرار باستثناءات:

Input: Training Dataset TD; Attribute List L; Splitting Criteria
Output: A decision tree T with Exceptions
Begin
(1) Generate root node N
(2) ***If*** each instance in TD falls in class Ck ***Then***
Return the leaf node N with class label Ck
End-if
(3) ***If*** L = Ø ***Then***// attribute list empty Treat N as leaf-node and label it with majority class label Cm and return
End-if
(4) Find the best splitting attribute (SA)based on the given criterion for attribute relevance
(5) Label this node with SA
(6) ***If*** SA is categorical and multi-way splits are allowed ***Then***
L ← L − SA; // delete splitting attribute from L
End-if
(7) ***For*** every outcome-value i of splitting attribute (SA)
perform data partition along the branches at the node and extend sub-tree for each partition
(7.1) Compute partition TDi
(7.2) ***If*** TDi= Ø ***Then*** // Empty partition
(a) Produce a leaf node ni and label it with majority class Cm
(b) Produce a leaf node ni' and label it with another class Cm'≠ Cm
Else
Create the node produced by decision tree (TDi, L) to node N
(7.3) Let node ni denotes rule Ri and node ni' denotes rule Ri'
(7.4) Generate a set TP (true positive examples) covered by the rule Ri
(7.5) Create a set FP (false positive examples) covered by the rule Ri'
(7.6) Compute $\gamma 1$, $\gamma 2$ and $\gamma 3$
(7.7) ***If***($\gamma 1 < 1$) ***Then***
If($\gamma 1 >> \gamma 2$) && ($\gamma 2 <$ tE) && ($\gamma 3 == 1$) ***Then***
Add Exceptions to node NE
Else
Add a normal node N
End if
End for
(8) Return N
End

الشّكل (٢) الخوارزمية المحسّنة لاكتشاف الاستثناءات

نلاحظ أن شجرة القرار قد أنشأت من بيانات العينة، وتُلحق الاستثناءات في نهاية كل عقدة طرفية، ولا تنتمي هذه الاستثناءات إلى فئة واحدة فقط. ولتوضيح ذلك سنأخذ شجرة القرار في الشّكل (٣) المبنية باستخدام الخوارزمية المحسّنة السابقة والتي تستخدم مجموعة بيانات الفطر كبيانات التدريب.



الشكل (٣) شجرة قرار لمجموعة بيانات الفطر

تعطى القواعد المستخلصة من الشجرة السابقة وفق الآتي:

*R*1: *If* (Odor= a) *Then* edible (4.0/0)
*R*2: *If* (Odor=c) *Then* poisonous (2.0/0)
*R*3: *If* (Odor=f) *Then* poisonous (6.0/0)
*R*4: *If* (Odor=l) *Then* edible (5.0/0)
*R*5: *If* (Odor=n) *Then* edible (10.0/2.0)
*R*6: *If* (Odor=p) *Then* poisonous (4.0/0)

تمثل الأرقام الموضحة بين القوسين عدد الحالات الإجمالية التي تغطيها فرضية القاعدة والمثيلات التي تنتمي إلى فئة الأقلية في عقدة التنفيذ. يوضح الجدول (٢) مصفوفة الارتباك (التي لا يمكننا أن نحدد فيها هل الفطر سام أم لا) الناتجة عن هذه القواعد. ويشير الرمزان P و D إلى بداية الشجرة وأجزاء القرار من القواعد.

الجدول (٢) مصفوفة الارتباك الخاصة بنموذج شجرة القرار

	المجموعة الصالحة للأكل المتوقعة		المجموعة السامة المتوقعة	
الفعلي صالح للأكل	(TP): $P \rightarrow D$	17	(FN): $\neg P \rightarrow D$	0
الفعلي سام	(FP): $P \rightarrow \neg D$	2	(TN): $\neg P \rightarrow \neg D$	12

بناءً على هذ الإجراء المتبع فإنه من الممكن أن تتأهل بعض الحالات التيُ صنّفت بشكل خاطئ عن طريق نموذج شجرة القرار المعتاد كاستثناءات صالحة لاستيعابها في شجرة القرار. بالنسبة لنموذج شجرة القرار أعلاه، فإن ١٧ حالة هي حالات إيجابية حقيقية True Positive (TP)، و١٢ حالة هي حالات سلبية حقيقية True Negative (TN)، ومثالان عبارة عن حالات إيجابية خاطئة False Positive (FP)، ولا تقع أي حالة في حالات سلبية خاطئة False Negative (FN). قد يكون هناك استثناءات غير ظاهرة في حالات FP أو FN. ففي المثال الذي أخذناه هناك حالات مخفية فقط في الوضع FP، ولمعرفة ما إذا كان هناك بعض الحالات في الوضع FP مؤهلة لأن تكون استثناء نحتاج إلى تحديد ثلاث متحولات إضافية هيγ_1 و γ_2 وγ_3 تعطى بالمعادلات (١و٢و٣) وفق الشروط الموضحة في (٤و٥و٦).

$\gamma_1 = TP/(TP + FP) = |P \wedge D|/|P|$ **(1)**

$\gamma_2 = FP/(TP + FP) = 1 - \gamma_1 = |P \wedge \neg D|/|P|$ **(2)**

$\gamma_3 = TN_E/(TP + FP \wedge E) = |P \wedge E \wedge \neg D|/(|P \wedge E|) = 1$ **(3)**

Where $\gamma_1 < 1$; **(4)**

$\gamma_1 > \gamma 2$ **(5)**

$\gamma_3 = 1$ **(6)**

يشير المتحولان γ_1 و γ_2 إلى نسب TP و FP. يستخدم الرمز E لتحديد زوج قيمة السمة الذي يُختبر ليكون مؤهلاً كاستثناء. يؤدي اكتشاف الاستثناء إلى تحويل بعض حالات FP إلى حالات TN، وتمثل TN_E مجموعات بيانات التدريب التي كانت تقع فيFP قبل اكتشاف الاستثناءات ولكنها أصبحت الآن حالات TN. يُراجع بعدها قرار القاعدة في وجود استثناء، والمثال الذي صُنّف بشكل خاطئ في وقت سابق ينتمي إلى فئة إيجابية لهذا صُنّف الآن بشكل صحيح على أنه ينتمي إلى فئة سلبية. يضمن المتحول الثالث γ_3 أن زوج قيمة الواصفة مؤهل كاستثناء فقط وفقط إذا حدث في حالات FP وليس في أي مكان من حالات TP لبيانات التدريب. بعبارة أخرى ستكون مقدمة القاعدة التي تُعزز مع الاستثناء دائمًا واحدة.

فعلى سبيل المثال، القاعدة R_5 هي المنافس الذي قد يكون له استثناءات. تحتوي هذه القاعدة على ٨ حالات في مجموعة TP وحالتين في مجموعة FP من إجمالي ٣١ حالة. قيمة γ_1 هي ٠.٨ وهي أقل من ١ وقيمة $\gamma 2$ هي ٠.٢ وهي أقل بكثير من ١. لهذا تُختبر الآن أزواج قيمة الواصفة التي لم تحدث بالفعل في الجزء الأساسي لمعرفة ما إذا كان أي من أزواج قيمة الواصفة هذه مؤهل لكي يكون استثناءات. لنقوم الآن بتعديل R_5 إلى R'_5 من خلال أن نلحق بها بعض الاستثناءات على النحو الآتي. وتُعرض شجرة القرار المحسّنة في الشكل (٤).

R'₅: *If* (odor = n) *Then* (decision = e) ***Unless*** (CapColor = p) ∨ (CapColor = y) ∨

(GillColor = r) ∨ (GillColor = y) (2.0): Poisonous

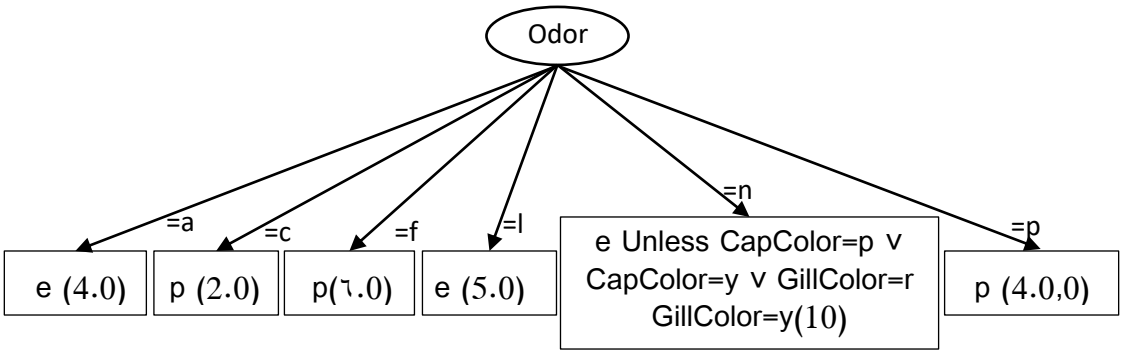
دعونا الآن نتحقق من قيم المتحولات $\gamma 1$ و$\gamma 2$ و$\gamma 3$:

$\gamma_1(R'_5) = TP/(TP + FP) = 8/10 = 0.8$

$\gamma_2(R'_5) = FP/(TP + FP) = 2/10 = 0.2$

$\gamma_3(R'_5('CapColor = p') = TN_E/(TP + FP \wedge E) = 1/1 = 1.0$

$\gamma_3(R'_5('GillColor = r') = TN_E/(TP + FP \wedge E) = 1/1 = 1.0$



الشّكل (٤) شجرة القرار المحسّنة بوجود الاستثناءات

تعمل القاعدة R_5 بوجود الاستثناءات على تغيير تسميات فئات الحالتين من "صالح للأكل" إلى "سامة"، وتُعد أمراً هاماً ويؤكد على أنه في معظم الحالات عندما تكون قيمة الواصفة "Odor" "رائحة" لا شيء (يُشار إليها بالحرف "n") عندها يُصنّف الفطر على أنه" صالح للأكل "، علماً أنه إذا كانت الواصفة عدم وجود رائحة صحيحة فإن هناك حالات استثنائية أخرى مثل "CapColor = p" أو "CapColor = y" أو "GillColor = r" أو "GillColor = y" وغيرها من الحالات الموجودة في القاعدة كما هو موضح في الشكل (٤)، يصبح صنف هذا الفطر ساماً. وبهذه الاستثناءات تتنبأ شجرة القرار بحالتين إضافيتين لمجموعة البيانات الفطر بشكل صحيح، وبالتالي يتم تحسين دقة المصنف أيضاً.

٤ التطبيق العملية للخوارزمية المحسّنة:

لاختبار فعالية الخوارزمية المحسّنة، أُستخدمت مجموعتان من البيانات هما مجموعة بيانات طلاب الجامعة الافتراضية السورية SVU التي يوجد فيها تخصصين خاصين بماجستير المعلوماتية هما: MWS (Master of Web Science) ماجستير دراسات عليا في علوم الويب والأخر هو MWT (Master of Web Technology) ماجستير التأهيل والتخصص في تقانات الويب، ومجموعة بيانات الفطر Mushroom. تحتوي مجموعة بيانات الطلاب على ١٠١ حالة بثلاثة واصفات وقيمتان للماجستير. وتحتوي مجموعة بيانات Mushroom على ٥٦٤٤ حالة و٢٤ واصفة مع قيمتين لتسميات الفئة. عُولجت مجموعات بيانات الفطر والتصويت مسبقًا وأُزيلت الحالات ذات القيم المفقودة من مجموعات البيانات هذه. يعرض الجدول(٣) نتائج مجموعات البيانات الثلاث. يعطي العمود الأول اسم مجموعات البيانات، بينما يسرد العمود الثاني القواعد المعززة بالاستثناءات التي أُكتشفت (إن وجدت) من خلال تطبيق الخوارزمية المحسّنة. تعرض بقية الأعمدة قيم المتحولات ودقة نماذج شجرة القرار بدون الاستثناءات ومعها. أُجريت جميع التجارب على نظام Window١٠ باستخدام لغة البرمجة JAVA. يحتوي الجدول(٣) على القواعد التي لها قيمة أقل من ١ بالنسبة لـ γ_1 و γ_2 وهي قواعد مرشحة أن يكون لها استثناءات. والاستثناءات هي أجزاء نادرة من المعرفة القيّمة والموجودة في أجزاء صغيرة من مجموعة البيانات، لهذا السبب تحظى بدعم منخفض.

الجدول (٣) نتائج مجموعات بيانات ماجستير الجامعة الافتراضية SVU والفطر Mushroom

Dataset	Rules	γ_1	γ_2	Accuracy without exception	ccuracy with Exception
MWS, MWT Master in SVU University	If (CertificateAvg >= 85 ∧ ExamDegree >= 85 ∧ BachelorAvg >= 95) Then (Decision = MWS_Master)	١.٠٠٠٠			
	If (CertificateAvg >= 85 ∧ ExamDegree >= 85 ∧ BachelorAvg < 95) Then (Decision = MWS_Master)	١.٠٠٠٠			
	If (CertificateAvg >= 85 ∧ ExamDegree < 85 ∧ BachelorAvg >= 95) Then (Decision = MWS_Master)	٠.٨٣٣	٠.١٦٧		
	If (CertificateAvg < 85 ∧ ExamDegree < 85 ∧ BachelorAvg >= 95) Then (Decision = MWS_Master)	١.٠٠٠٠			
	If (CertificateAvg >= 75 ∧ ExamDegree >= 75 ∧ BachelorAvg >= 90) Then (Decision = MWT_Master)	١.٠٠٠٠		92.08%	100%
	If (CertificateAvg >= 75 ∧ ExamDegree >= 75 ∧ BachelorAvg < 90) Then (Decision = MWT_Master)	١.٠٠٠٠			
	If (CertificateAvg >= 75 ∧ ExamDegree < 75 ∧ BachelorAvg >= 90) Then (Decision = MWT_Master)	١.٠٠٠٠			
	If (CertificateAvg < 75 ∧ ExamDegree < 75 ∧ BachelorAvg >= 90) Then (Decision = MWT_Master)	١.٠٠٠٠			
Mushroom	*If* (odor = a) *Then* (*Decision* = e)	١.٠٠٠			
	If (odor = c) *Then* (*Decision* = p)	١.٠٠٠			
	If (odor = f) *Then* (*Decision* = p)	١.٠٠٠			
	If (odor = l) *Then* (*Decision* = e)	١.٠٠٠			
	If (odor = m) *Then* (*Decision* = p)	١.٠٠٠			
	If (odor = n ∧ spore-print-color = b) Then (Decision = e)	١.٠٠٠			
	If (odor = n ∧ spore-print-color = h) Then (Decision = e)	١.٠٠٠			
	If (odor = n ∧ spore-print-color = k)	١.٠٠٠			

	Then (Decision =e)				
	If (odor = n ∧ spore-print-color = n) Then (Decision = e)	١.٠٠٠			
	If (odor = n ∧ spore-print-color = o) Then (Decision = e)	١.٠٠٠			
	If (odor = n) Then (Decision = e) Unless (spore-print-color = r)(Decision = p)	٠.٩٨٠	٠.٠٢٠	98.52%	100%
	If (odor = n ∧ spore-print-color = w ∧ gill-size = b) Then (Decision = e) If (odor = n) Then (Decision = e) Unless (spore-print-color = w ∨ gill-spacing = c) (Decision = p)	١.٠٠٠			
	If (odor = n) Then (Decision = e) Unless (spore-print-color = w ∧ gill-spacing = w ∧ population = c) (Decision = p)	٠.٩٢٢	٠.٠٧٨		
	If (odor = n ∧ spore-print-color = w ∧ gill-spacing = w ∧ population = v) Then (Decision = e)	٠.٩٤٧	٠.٠٢٦		
	If (odor = n ∧ spore-print-color = y) Then (Decision = e)	١.٠٠٠			
	If (odor = p) Then (Decision = p)	١.٠٠٠			

نلاحظ أنه يمكن التحسين في الدقة بشكل كبير في حال كانت الاستثناءات موجودة في العديد من البيانات الصغيرة المنفصلة من خلال تقليل درجة تعقيد الخوارزمية. وتشير القواعد مع الاستثناءات إلى حالات شاذة مهمة تحتوي على الكثير من المعرفة المثيرة للاهتمام، تمنع المصنّف من القيام بإجراءات مسبقة خاطئة في ظروف استثنائية نادرة.

٥. الخلاصة والتوصيات:

أُقترح في هذا البحث خوارزمية محسّنة لشجرة القرار والتي تستوعب الاستثناءات المخفية في مجموعات البيانات الصغيرة في نموذج شجرة القرار. ويمتاز النموذج فيها بأنه يحتوي على قواعد تراجع الاستنتاجات في ظل ظروف نادرة واستثنائية. كما أن مصنّف شجرة القرار لا يفيد فقط بأنه أكثر دقة بل يُعدُّ أيضاً مختصراً وفيه معرفة مهمة بشكل كبير. إن معرفة القواعد العامة لتصنيف الصف أمراً ضرورياً، إلا أنه وخلال استكشاف ومعرفة الاستثناءات، يصبح المصنف ذكياً بدرجة كافية لتجنب التنبؤات غير الصحيحة في الظروف الاستثنائية. ويمكن أن يستفاد من هذا المصنّف في تطبيقات مختلفة كالتطبيقات في مجال الروبوتات، واكتشاف الاحتيال، والتشخيص الطبي الخاطئ. ويمكن مستقبلاً توسيع هذا البحث لاستيعاب كافة الاستثناءات الموجودة في مجموعة البيانات.

٦. المراجع:

[1].Sunil Kumar1, Saroj Ratnoo1, Renu Bala: *Pattern Recognition Letters*. Haryana, India. 28, 7 (2016) 825–832

[2].Appavu alias Balamurugan, S., Rajaram, R.*: Effective Solution for Unhandled Exception in Decision Tree Induction Algorithms.* Expert Systems with Applications. 36, 10, (2009) 12113–12119

[3].Bala, R., & Saroj.: *Discovering Fuzzy Censored Classification Rules (FCCRs): A Genetic Algorithm Approach*. International Journal of Artificial Intelligence & Applications, 3, 4 (2012) 175–188

[4].Bala, R., Ratnoo, S.: *A Genetic Algorithm Approach for Discovering Tuned Fuzzy Classification Rules with Intra- and Inter-Class Exceptions*. Journal of Intelligent Systems. 25 (2016) 263–282

[5]. Bharadwaj, K.K., Al-Maqaleh, B.M.: *Evolutionary Approach for Automated Discovery of Censored Production Rules*. 1, 10 (2007) 3230-3235

[6]. Compton, P. et al.*: Ripple Down Rules: Turning Knowledge Acquisition into Knowledge Maintenanc*e. Artificial Intelligent Medicine. 4, 6 (1992) 463–475

[7].Carvalho, D.R., Freitas, A.A.*: A Hybrid Decision Tree/Genetic Algorithm Method for Data Mining. Information Sciences*. 163, 1 (2004) 13–35

[8].Dietterich, T.G.*: Ensemble Methods in Machine Learning. In: Multiple Classifier Systems*. Springer Berlin Heidelberg (2000) 1–15

[9].Gehrke, J. et al.*: RainForest- A Framework for Fast Decision Tree Construction of Large Datasets.* Data Mining and Knowledge Discovery. 4, 2-3 (2000) 127–162

[10]. Geng, L., Hamilton, H.J.*: Interestingness Measures for Data Mining*: A Survey. ACM Computing Surveys. 38, 3, 9 (2006).

[11]. Kamiński, B.*; Jakubczyk, M.; Szufel, P:*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5767274 . European Journal Research. (2017) 135–159.

[12]. Barsacchi, M.; Bechini, A.; Marcelloni, F: *An analysis of boosted ensembles of binary fuzzy decision trees*". Expert Systems with Applications. (2021) 113-154.