الفصل الخامس

التحليل التمييزي النوعي (Qualitive)

(شجرة التصنيف والانحدار CART)

5-1 : تمهيد:

يتناول التحليل التمييزي النوعي دراسة تأثير المتحولات النوعية أو المختلطة (الكمية والنوعية) على تصنيف عناصر المجتمع المدروس إلى مجموعات منفصلة (متجانسة أو نقية)، ويعتمد هذا التحليل على تصميم شجرة تشبه شجرة القرارات تسمى شجرة التصنيف والانحدار (CART)، المأخوذة من الكلمات (Classification And Regression Tree)، وينطلق هذا التحليل من اعتبار أن جميع عناصر المجتمع المدروس تشكل مجموعة واحدة، ونريد تصنيفها إلى مجموعتين عمليتين أو أكثر، ثم تفريع هاتين المجموعتين إلى مجموعتين جديدتين أو أكثر، وتتم عملية التفريع واتخاذ القرارات من خلال فواصل توضع على مفاصل الشجرة، وتعطينا عند كل مفصل (عقدة) فرعين أو أكثر، وتستمر عملية التفريع حتى يتحقق أمر التوقف الذي يحدده الباحث .

وتبدأ شجرة التصنيف من جذر واحد (يسمى بالعقدة الأصلية أو عقدة الأب Parent)، ويتفرع عنه أغصان على شكل أسهم لتتصل بعقد فرعية أخرى تسمى بالعقد الداخلية Internal node (أو بعقد الأبناء والأحفاد)، ثم تنتهي بعقد ختامية تسمى بالعقد الخارجية Terminal Node، وهي تضم نتائج التصنيف النهائية ويطلق عليها أيضاً اسم الورقة (laef) وهي ثمرة التصنيف .

وترسم شجرة التصنيف بصورة مقلوبة، وتمثل العقدة الأصلية والعقد الداخلية بدوائر (أو بقطوع ناقصة) ويكتب بداخلها شروط التصنيف أو التفريع. أما العقد الخارجية فتمثل على شكل مربعات وتتضمن نتائج التصنيف وتضم عناصر (واحد على الأقل) مفروزة من المجتمع المدروس (انظر الشكل 5-1)، وتتم عملية الفرز في كل عقدة حسب الجواب على السؤال الذي فيها. فإذا كان السؤال ثنائي الجواب (نعم أو لا) وكان الجواب بـ (نعم) فإننا نضع (yes) على السهم اليساري، أما إذا كان الجواب بـ (لا) فإننا نضع (No) على السهم اليميني .

ولكن إذا كان جواب السؤال متعدد الحالات فإننا نعمل على وضعها في نظام معين في كل العقد الداخلية التي تتضمن مثل ذلك السؤال .

مثال (5−1): لنفترض إننا نريد تصنيف عينة من موظفي الجامعة (حجمها $n = 100$) حسب ثلاثة متحولات هي: الجنس− الحالة الزوجية− حالة المسكن. إلى مجموعات متجانسة من حيث المسكن، فوجهنا إليهم ثلاثة أسئلة مع أجوبتها الممكنة، وطلبنا من كل منهم تحديد الجواب الذي ينطبق عليه، وكانت الأسئلة كما يلي:

س1: الجنس : | ذكر | أنثى |

س2: الحالة الزواجية: | متزوج | غير متزوج | (بدون تفاصيل)

س3: حالة السكن: | ملك | إيجار | عند الأهل أو الأقارب |

ولنفترض أن تكرارات الإجابات كانت كما في الجدول التالي:

جدول (5−1) تصنيف أفراد العينة انطلاقاً من حالة الجنس ثم الحالة الزواجية ثم حالة المسكن:

الجنس	العدد	الحالة الزوجية	العدد	حالة السكن	العدد
ذكر	60	متزوج	45	ملك	20
				إيجار	15
				عند الأهل	10
		غير متزوج	15	ملك	5
				إيجار	3
				عند الأهل	7
أنثى	40	متزوجة	12	ملك	7
				إيجار	3
				عند الأهل	2
		غير متزوجة	28	ملك	2
				إيجار	1
				عند الأهل	25
المجموع	100	المجموع	100	المجموع	100

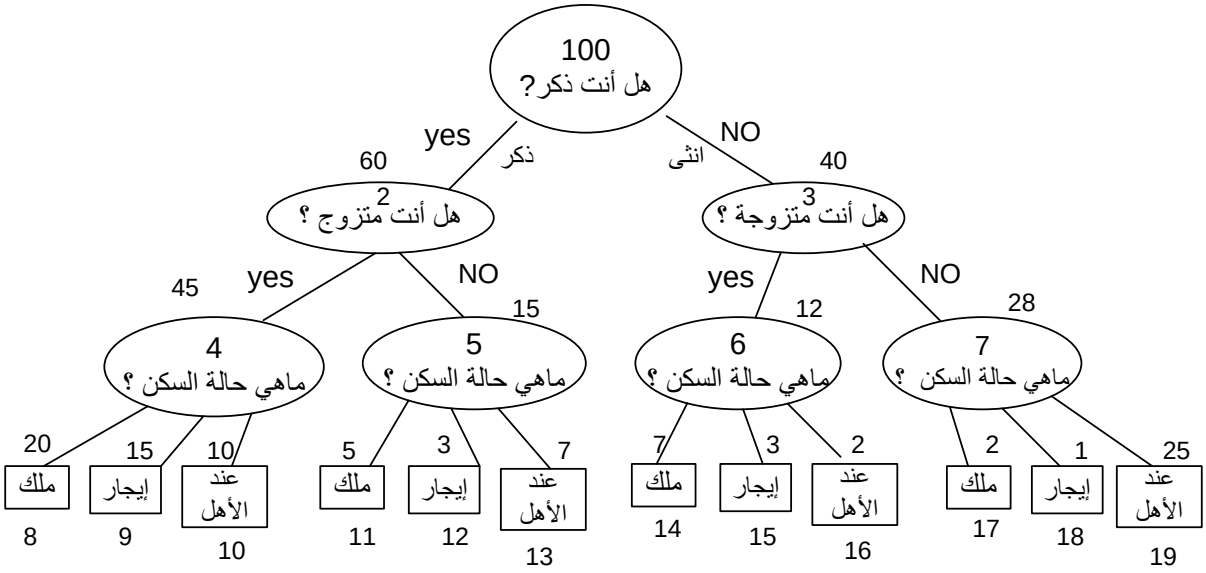
ومن الجدول السابق نلاحظ أنه يمكننا فرز أو تصنيف أفراد هذه العينة إلى مجموعات متجانسة وبأشكال مختلفة. وهنا بدأت عملية فرز هذه المجموعات إلى مجموعتي الجنس (ذكور اناث)، ثم إلى مجموعتي المتزوجين وغير المتزوجين (من كل جنس ومن الجنسين)، ثم إلى مجموعات حالات السكن (ملك، إيجار، عند الأهل)، وذلك حسب كل حالة من الحالات السابقة . كما نلاحظ أنه يمكننا استخلاص وتشكيل المجموعات المتجانسة من حيث المسكن كما يلي:

G_1 مجموعة المالكين من جميع الحالات وعدد عناصرها يساوي: $N_1 = 20 + 5 + 7 + 2 = 34$

G_2 مجموعة المستأجرين من جميع الحالات وعدد عناصرها يساوي: $N_2 = 15 + 3 + 3 + 1 = 22$

G_3 **مجموعة الساكنين عند الأهل لجميع الحالات وعدد عناصرها أنه يساوي:** $N_3 = 10 + 7 + 2 + 25 = 44$

وأخيراً يمكننا رسم شجرة التصنيف لهذه العينة كما يلي:



الشكل (5-1): شجرة التصنيف لأفراد العينة المدروسة

كما يمكننا تشطير الشجرة إلى شجيرات فرعية بشرط أن يكون لكل شجيرة فرعية جذر واحد، أي أن يكون على رأس كل شجيرة فرعية أصل واحد، فمثلاً يمكننا أن نشطر هذه الشجرة إلى شجرتين فرعيتين هما: شجيرة الذكور وأصلها في العقدة (2)، وشجيرة الإناث وأصلهما، في العقدة (3). وكذلك يمكننا اعتبار المتزوجين الذكور شجيرة فرعية أصلها العقدة (4)، كما يمكننا اعتبار الإناث المتزوجات شجيرة فرعية أصلها في العقدة (6).

وأخيراً نشير إلى أن هذا التصنيف يقسم فضاء العينة إلى (12) مجموعة متجانسة هي عبارة عن مجموعات العقد الخارجية كما هو مبين على الشكل التالي:

الاناث		**الذكور**	
غير المتزوجات	**المتزوجات**	**غير متزوج**	**متزوج**
المجموعة 17	**المجموعة** 14	**المجموعة** 11	**المجموعة** 8
المجموعة 18	**المجموعة** 15	**المجموعة** 12	**المجموعة** 9
المجموعة 19	**المجموعة** 16	**المجموعة** 13	**المجموعة** 10

الشكل (5-2): فضاء العينة المقسم

5-2 : مراحل تصميم شجرة التصنيف :

لقد لاحظنا من المثال السابق أن شجرة التصنيف تأخذ شكلاً هرمياً، يبدأ من عقدة واحدة تسمى عقدة الأصل (الجذر)، ويتفرع عنها عبر الأغصان عقد داخلية متعددة وذات مستويات مختلفة، وينتهي التفرع بما يسمى بالعقد الخارجية، وإن حجم أو عمق الشجرة يتوقف على أهداف البحث أو على الشروط التي يضعها الباحث على عملية التصنيف .

إن عملية تصميم شجرة التصنيف تمر بعدة مراحل هي كما يلي:

1- مرحلة الإنشاء أو البناء وتشمل هذه المرحلة عمليات النمو وأهمها عمليات التفرع أو الانشطار Splitting وتتألف من الخطوات التالية:

أ- تحديد المتحول التابع Y والمؤلف من المجموعات التصنيفية $Y = G_1, G_2, \dots G_g$.

ب- اختيار المتحولات التصنيفية أو التفسيرية وتحديد تسلسل تطبيقها عند كل عقدة، بما في ذلك عقدة الأصل. وإذا كان عدد المتحولات كبيراً فإنه يجب اختصارها باختيار المتحولات الهامة منها وحذف المتحولات غير الهامة منها، وذلك وفق معايير محددة. ولنرمز لهذه المتحولات بـ $X_1, X_2, X_3 \dots X_p$.

ج- اختيار عقدة الأصل بحيث تكون مناسبة لأهداف البحث، ومتوافقة مع معايير الانشطار المتسلسلة والممعيرة لاستنباط الأبناء والأحفاد من ذلك الأصل، وهنا نشير إلى أنه يمكننا الحصول على أشجار متعددة إذا غيرنا الجذور المعتمدة في عقدة الأصل .

د- تحديد قواعد التفرع (الانشطار) عند كل عقدة t ابتداء من عقدة الأصل. وتحديد الحدود العددية أو الفئات النوعية لكل متحول X_i عند كل عقدة t .

فإذا كان المتحول X_i كمياً فإنه يمكننا أن نضع قاعدة التفرع على شكل متراجحات كما يلي:

$$X_i \geq C_1 \quad \text{أو} \quad X_k \geq C_k \quad \text{أو} \quad X_i + X_k \geq C_3 \qquad (1-5)$$

حيث C_3, C_k, C_1 أعداد حقيقية من مجال تحول X_i , X_k .

أما إذا كان المتحول X_i نوعياً فإنه يمكننا أن نضع قاعدة التفرع على شكل إشارة انتماء كما يلي:

$$X_i \in G_j \quad \text{أو} \quad X \notin G_j \qquad (2-5)$$

حيث أن: G_j هي احدى مجموعات تابع المخرجات النوعي Y، والمؤلف من المجموعات المحددة التالية :

$$Y = \{G_1, G_2, \dots G_g\}$$

وهنا يجب أن نشير إلى أنه يجب صياغة واختيار قاعدة التفرع بحيث تكون البيانات الناتجة عن عملية التفرع في العقد التالية متجانسة وأكثر نقاوة من البيانات التي كانت في العقدة الأولى . وهناك معايير خاصة لاختيار قاعدة التفرع المناسبة سنعرضها لاحقاً في فقرة خاصة .

2- عملية التقسيم أو التجزئة (Partition): وفيها تتم عملية تقسيم البيانات X في كل عقدة إلى مجموعتين أساسيتين منفصلتين (أو أكثر) حسب قاعدة التفرع ، وبذلك يتم تقسيم الفضاء R^P إلى قسمين (أو أكثر)، لكل منها سمات خاصة. ثم نقوم بتكرار ذلك التقسيم مرة أخرى فنحصل على مجموعتين منفصلتين جديدتين مقابل كل مجموعة سابقة .

وهكذا نكرر عمليات التقسيم حتى تتحقق القاعدة المخصصة للتوقف عن التقسيم .

3- عملية التوقف (Stoping) وتتم حسب قاعدة معينة يضعها الباحث حسب طبيعة وهدف البحث، ويمكن أن تكون من أحد الأشكال التالية:

- إذا أصبح التغير في قيمة تابع الخطأ (الشوائبية) صغير جداً أو أقل من الحد المفروض عليه .
- إذا أصبح عدد العناصر عند التقسيم الأخير في إحدى المجموعات الناتجة عنه صغيراً (أقل من 5) أو أصبح يساوي الواحد . أو أصبحت نسبتهم صغيرة بالنسبة للمجموعات الناتجة الأخرى .
- إذا أصبح حجم الشجرة أو عمقها كبيراً (يحدده الباحث) والمقصود بعمقها عدد مستويات التفرع فيها. أما حجم الشجرة فيقاس بعد العقد الخارجية فيها .
- إذا أصبح مستوى الدقة محققاً أو أصبحت المجموعات الناتجة نقية تماماً .

4- عملية التقليم أو التشذيب (Pruning): وهي عبارة عن إسقاط بعض الفروع السابقة، الصادرة عن بعض العقد الداخلية (مع العقد الخارجية التابعة لها). والإبقاء على الفروع الضرورية اللازمة لأغراض البحث. ولكن ذلك يجب أن يتم وفق قواعد محددة سنتعرض لها لاحقاً .

5- التجميع (Grouping) وهو عبارة عن تجميع العقد الخارجية النهائية في مجموعات مناسبة لكل منها. بحيث يكون معدل التصنيف الخاطئ فيها أصغر ما يمكن .

6- رسم الشجرة وترقيم العقد الداخلية والخارجية ترقيماً تصاعدياً حسب مستويات التفرع، وتبدأ عملية الترقيم بوضع الرقم $t = 1$ للعقدة الأصلية (عقدة الجذر)، ثم إعطاء العقد الناتجة عنها أرقاماً متتالية متصاعدة مثل (2) و(3) و(4) و(5)، وتوضع أرقام العقد ضمن دائرتها. وأحياناً توضع أرقام العقد الخارجية تحتها. ويفضل أن تتم عملية الترقيم حسب مستويات العمق ومن اليسار إلى اليمين كما هو موضح على الشكل (5-1). فإذا كان رقم العقدة الداخلية t، فإن رقم العقدة الناتجة عنها يمكن أن يكون أي رقم t' أكبر من الرقم t، أي يجب أن يكون $t' > t$.

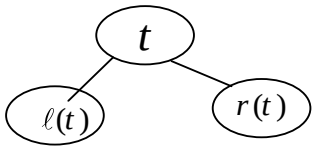
5-3 : كيفية تصميم شجرة التصنيف (حسب التفرع الثنائي Binary)

لنفترض أنه لدينا شجرة T مؤلفة من جملة من العقد الداخلية والخارجية، والمرقمة تصاعدياً بأرقام صحيحة موجبة وحسب مستويات التفرع، وتحقق العلاقة $t' > t$، حيث t' هو رقم العقدة المتفرعة عن العقدة t، فإذا كان الرقم $t = 1$ للعقدة الأصلية (عقدة الجذر) فإن الأرقام (2) و(3) أو(4) أو(5) ستكون للعقد الناتجة عنها .

ولنعرف الآن على كل عقدة t، تابعين $\ell(t)$ و r(t) كما يلي:

$\ell(t)$ - تابع يعبر عن رقم العقدة اليسارية الناتجة عن التفرع عند النقطة t (من كلمة left) .

r(t)- تابع يعبر عن رقم العقدة اليمينية الناتجة عن التفرع عند النقطة t (من كلمة right) .



بحيث يحقق هذان التابعان في حالة التفرع الثنائي الخواص التالية:

1- من أجل أية عقدة $t \in T$ فإن $\ell(t) > 0$ و $r(t) > 0$.

وإذا كانت العقدة الناتجة عن t عقدة داخلية هي t' فإنه يكون لدينا

$\ell(t) > t$ و $r(t) > t$.

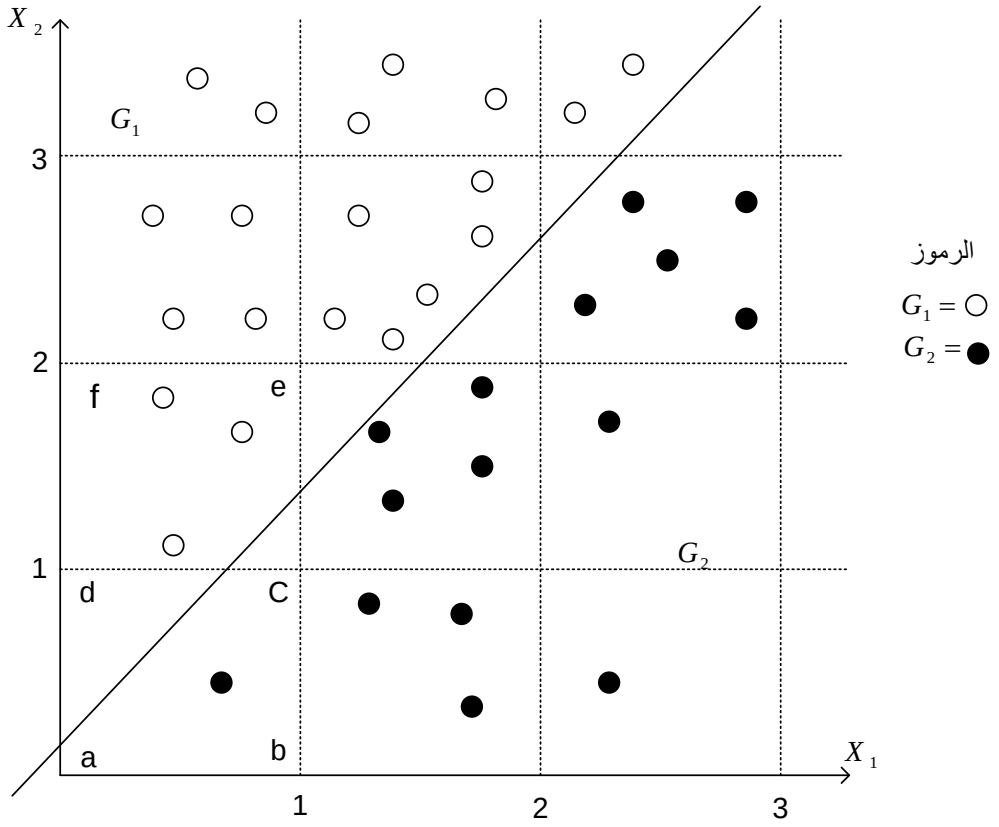
أما إذا كانت العقدة t' خارجية (ورقة) فإننا نجعل هذين التابعين يأخذان قيمة الصفر، أي نجعلهما يساويان $\ell(t) = 0$ و $r(t) = 0$. وذلك للدلالة على عدم وجود عقد بعد الأوراق .

2- لكل عقدة t من الشجرة T (ماعدا عقدة الأصل حيث $t = 1$) يوجد أصل وحيد S من T، ولا توجد عقدة بدون أصل، أي أنه يوجد أصل وحيد s لكل عقدة t $(t \neq 1)$ بحيث يكون: $t = \ell(s)$ أو $t = r(s)$.

مثال (5-2): لنفترض إننا نريد تصنيف (35) عنصراً إلى مجموعتين متجانستين، حسب تغيرات متحولين كميين X_1 و X_2 معرفين في المستوى على أن: $X_1 \geq 0$ و $X_2 \geq 0$، علماً بأن هذه العناصر تؤلف قبل التصنيف مجموعتين G_1 و G_2 فيهما (20) و (15) عنصراً على الترتيب .

ولنفترض أنه بعد الدراسة والرسم تبين لنا أن تلك العناصر تتوزع على المستوى $X_1 0 X_2$ كما يلي:

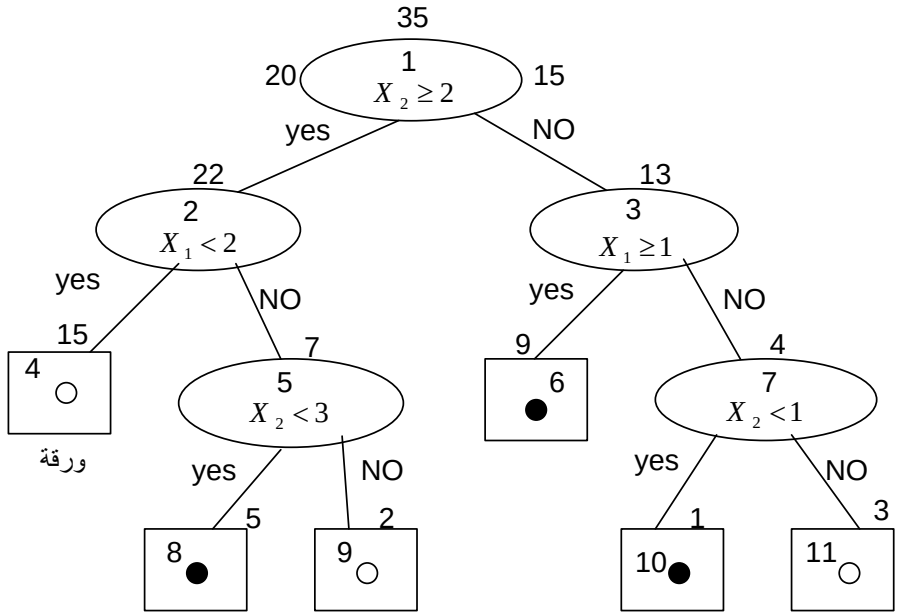
(المصدر: *Webb. A. R p.227* بتصرف)



الشكل (5-3) التوزيع البياني لعناصر العينة : $G_1 = \bigcirc$ و $G_2 = \bullet$

ولنفترض أن الباحث قد قام بتصميم شجرة التصنيف انطلاقاً من الشرط $(X_2 \geq 2)$ كمايلي:

t	$\ell(t)$	r(t)
1	2	3
2	4	5
3	6	7
4	0	0
5	8	9
6	0	0
7	10	11
8	0	0
9	0	0
10	0	0
11	0	0



الشكل (5−4): شجرة التصنيف للمثال (5−2) : $G_1 = \bigcirc$ و $G_2 = \bullet$

والآن نعرف التابعين $\ell(t)$ و r(t) على الشجرة المرسومة على الشكل (5−4)، فنجد أن قيمتهما عند كل نقطة t تساويان كما في الجدول المرافق لذلك الشكل . ولقد أشرنا سابقاً إلى أن هذه التفرعات تقسم الربع الموجب للفضاء R^2 حسب رقم المتحولين X_1 وX_2 إلى مناطق منفصلة، ونحصل عليهما بشكل متتالٍ كما يلي:

بما أن قاعدة التفرع في العقدة الأولى الأصلية هي: $(X_2 \geq 2)$ فهي تقسم الربع الموجب إلى قسمين: قسم تحت الخط $(X_2 = 2)$ وقسم فوقه، وضمن هذا التقسيم نأخذ الفرع الأيمن من الشجرة حيث الجواب (No) والموافق لـ $(X_2 < 2)$، فنجد أن قاعدة التفرع في العقدة (3) هي: $(X_1 \geq 1)$ وهي تعطينا بتقاطعها مع $(X_2 < 2)$ (حيث الفرع (No)) منطقتين تتحددان كما يلي:

إذا كان الجواب في العقدة (3) على $(X_1 \geq 1)$ بـ (yes)، فإننا نحصل على المنطقة التي نقع تحت المستقيم $(X_2 = 2)$ وعلى يمين المستقيم $(X_1 = 1)$، وهي تقابل العقدة الداخلية (6) لذلك رمزنا لها على الشكل (5−5) بـ R(6) وهي تضم (9) عناصر من G_2 فقط .

أما إذا كان الجواب على $(X_1 \geq 1)$ بـ (No)، فإننا نحصل على المنطقة الواقعة تحت المستقيم $(X_2 = 2)$ وعلى يسار المستقيم $(X_1 = 1)$، وهي تشمل المستطيل $a\ b\ e\ f$ المقابل للعقدة الداخلية (7) .

وضمن هذا المستطيل نختبر نتيجة قاعدة التفرع في العقدة (7) والتي هي: $(X_2 < 1)$ فنجد أنه: إذا كان الجواب بـ (yes)، فإننا نحصل على المستطيل $a\ b\ c\ d$، والذي يضم عنصراً واحداً من G_1. وهو يقابل العقدة الخارجية (10)، ولذلك رمزنا لها على الشكل (5−5) بـR(10)، أما إذا كان الجواب بـ (No) فإننا نحصل على المستطيل $d\ c\ e\ f$ ورمزنا له على الشكل (5−5) بـ R(11)، لأنه يقابل العقدة الخارجية (11)، وهي تضم (3) عناصر من G_1 .

والآن لنذهب إلى الفرع الأيسر حيث الجواب بـ (yes) على القاعدة الأصلية $(X_2 \geq 2)$، فنجد أن قاعدة التفرع في العقدة (2) هي: $(X_1 < 2)$، فإذا كان الجواب بـ (yes) فإننا نحصل على المنطقة المفتوحة فوق $(X_2 = 2)$ وعلى يسار $(X_1 = 2)$، وهي تقابل العقدة الخارجية (4) لذلك نرمز لها بـ R(4) على الشكل (5-5)، وهي تضم (17) عنصراً من G_1 .

أما إذا كان الجواب بـ (No) فإننا نحصل على منطقة مفتوحة تقع فوق $(X_2 = 2)$، وعلى يمين المستقيم $(X_1 = 2)$، وهي تقابل العقدة الداخلية (5) .

وأخيراً نقوم باختيار نتيجة قاعدة التفرع في العقدة (5) والتي هي $(X_2 < 3)$، وندرس تقاطعها مع نتيجتي القاعدتين السابقتين لها وهما: (No) للقاعدة $(X_1 < 2)$ و(yes) للقاعدة $(X_2 \geq 2)$ فنجد أنه: إذا كان الجواب على $(X_2 < 3)$ بـ (yes) فإننا نحصل على المنطقة المفتوحة الواقعة فوق $(X_2 = 2)$ وتحت المستقيم $(X_2 = 3)$ وعلى يمين $(X_1 = 2)$، وهي تقابل العقدة الخارجية (8) لذلك نرمز لها بـ R(8) على الشكل (5-5)، وهي تضم (5) عنصراً من G_2 .

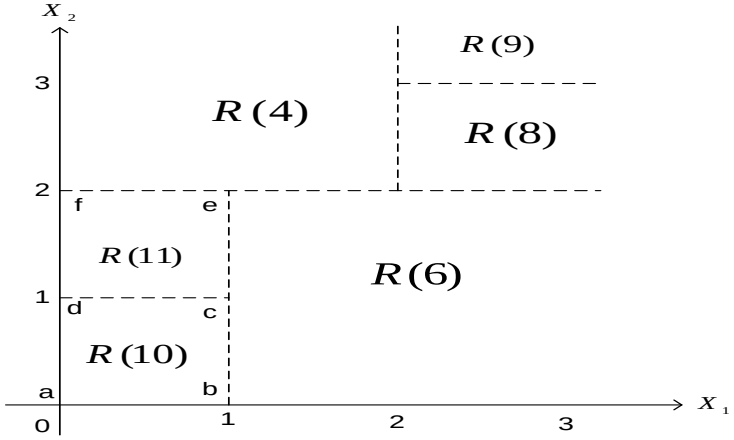
أما إذا كان الجواب على $(X_2 < 3)$ بـ (No) فإننا نحصل على المنطقة المفتوحة الواقعة فوق $(X_2 = 3)$، وعلى يمين المستقيم $(X_1 = 2)$، وهي تقابل العقدة الخارجية (9) لذلك نرمز لها بـ R(9) على الشكل (5-5)، وهي تضم عنصرين من G_1 .

وهكذا نجد أن عملية التصنيف على الشجرة أنجزت عمليتين بآن واحدهما:

1- تقسيم منطقة تعريف المتحولات X إلى مناطق منفصلة تقابل العقد الخارجية وهي: $R(4), R(6), R(8), R(9), R(10), R(11)$.

2- توزيع عناصر العينة على مجموعات مناسبة حسب إحداثياتها ومواقعها في المناطق السابقة وسندرسها لاحقاً في فقرة خاصة .

وأخيراً نرسم هذه المناطق كما يلي :



الشكل (5-5): المناطق المنفصلة لفضاء العينة

وأخيراً يمكننا تجميع هذه المناطق ضمن المجموعتين المفروضتين كما يلي:

المجموعة الأولى G_1 : وتضم المناطق $R(11), R(4), R(9)$ ونكتبها كما يلي:

$$G_1 = R(11) \cup R(4) \cup R(9) \qquad (3-5)$$

المجموعة الثانية G_2 : وتضم المناطق $R(8), R(6), R(10)$ ونكتب ذلك كما يلي :

$$G_2 = R(10) \cup R(6) \cup R(8) \qquad (4-5)$$

ملاحظة 1: عند تصميم الشجرة يجب الانتباه إلى عدم تعارض الشروط الواردة في قواعد التفرع (ضمن العقد الداخلية) مع بعضها البعض وخاصة على المسار الواحد .

فمثلاً لا يجوز أن تكون قاعدة الفصل في العقدة (5) متعارضة مع قاعدة الفصل في (2) أو في (1) وكأن نضع فيها الشرط $(X_2 < 2)$.

ملاحظة 2: إن المناطق $R(t)$ التي حصلنا عليها غير متقاطعة ومتكاملة أي أنها تحقق العلاقتين :

$$R(t) \cap R(s) = \phi = \quad \left(\text{مجموعة خالية}\right) \qquad (5-5)$$

$$\bigcup_{t=1}^{|T|} R(t) = \Omega \qquad : \ R^P \text{ فضاء العينة في} \qquad (5-6)$$

حيث $|T|$: هو عدد العقد الخارجية ويسمى حجم الشجرة .

ملاحظة 3: يمكن رسم الشجرة بطرائق مختلفة، وذلك بوضع نفس قواعد التفرع بطرائق مختلفة، فنحصل على شجرات جديدة تقسم فضاء العينة بطرائق مختلفة، ولهذا كان لابد من اختيار طريقة مثالية للتصنيف ولتوزيع قواعد التفرع على العقد الداخلية، بحيث يعطينا التقسيم الناتج مناطق نقية أو صافية، أي بحيث تضم كل منطقة عناصر من إحدى المجموعات فقط ، أو على مناطق تتضمن أقل عدد من العناصر الغريبة فيها (ارسم الشجرة السابقة بطريقة أخرى) .

5-4 : التصنيف حسب احتمالات الانتماء والتوزيع :

بما أن تصميم شجرة التصنيف يتم باستخدام خواص مجموعة البيانات المأخوذة من عناصر عينة حجمها n، والتي تتضمن قيم المتحولات X فيها، وقيم مستويات المجموعات المقابلة لها في التابع Y، فإنه يكون لدينا n مجموعة من القياسات المتقابلة $G_i \ , X_{1i} \ X_{2i} \dots X_{pi}$، حيث أن i هي دليل عناصر العينة ويأخذ $(i: 1\ 2\ 3 \dots n)$، وحيث أن n هو حجم العينة المدروسة، ويتم تنظيم ذلك في جدول البيانات الأولية المتقابلة .

والآن لنرمز بـ $N(t)$ لعدد عناصر العينة التي تنتمي إلى المنطقة $R(t)$ المقابلة للعقدة t، وهو يمكن أن يكون موزعاً على عدد من المجموعات G_j التي يتألف منها التابع Y .

ولنرمز بـ $N_j(t)$ لعدد عناصر العينة التي تنتمي إلى المنطقة $R(t)$ وإلى المجموعة G_j معاً، والتي تحقق العلاقة: $\sum_{j=1}^{g} N_j(t) = N(t)$، وعندها نجد أنه يمكننا حساب احتمالات الانتماء والتوزيع كمايلي:

- إن احتمال انتماء أي عنصر i من العينة n عند العقدة t إلى المنطقة $R(t)$ يساوي :

$$P(t) = \frac{N(t)}{n} \qquad (7-5)$$

- إن احتمال انتماء أي عنصر i من المنطقة $R(t)$ المقابلة للعقدة t، إلى المجموعة G_j يساوي :

$$P\big(G_j/i \in R(t)\big) = \frac{N_j(t)}{N(t)} \qquad (8-5)$$

- إن احتمال أن يتم توزيع عناصر العينة عند العقدة t، إلى عقدة اليسار $\ell(t)$ وإلى عقدة اليمين $r(t)$ يساويان :

$$P[\ell(t)] = \frac{N[\ell(t)]}{n} \qquad (9-5)$$

$$P[r(t)] = \frac{N[r(t)]}{n} \qquad (10-5)$$

- إن احتمال أن يذهب أي عنصر i متواجد في العقدة t، إلى إحدى عقدتي اليسار أو اليمين يساوي :

$$P_\ell = \frac{P[\ell(t)]}{P(t)} = \frac{N[\ell(t)]}{N(t)} \qquad (11-5) \quad \text{إلى العقدة اليسارية :}$$

$$P_r = \frac{P[r(t)]}{P(t)} = \frac{N[r(t)]}{N(t)} \qquad (12-5) \quad \text{إلى العقدة اليمينية :}$$

وهكذا يمكننا أن نحدد ملامح كل مجموعة G_j في كل عقدة t، وذلك حسب تناسبها مع عدد عناصر العينة التي تنتمي إليها من المنطقة $R(t)$ ، أي حسب احتمالاتها فيها، ثم نقارن الاحتمالات المقابلة لهذه المجموعات ونختار المجموعة G_k التي تقابل أكبر الاحتمالات .

القاعدة: نصنف العنصر i من العقدة t إلى المجموعة G_k إذا كان الاحتمال الشرطي المقابل لها أكبر من الاحتمالات الشرطية المقابلة للمجموعات الأخرى، ونكتب ذلك كما يلي. إذا كان:

$$P(G_k/t) = \max_{j=1}^{g} P\big(G_j/t\big) \qquad (13-5)$$

فإننا نصنف العنصر i من العقدة t في المجموعة G_k .

مثال (5-3): لنأخذ المثال (5-2) السابق ولنفترض أن عناصر العينة $(n = 35)$ مؤلفة من مجموعتين $(n_1 = 20)$, $(n_2 = 15)$ وتتوزع على العقدتين اللاحقتين $\ell(t)$, $r(t)$. ثم نقوم بحساب الاحتمالات السابقة حسب العلاقات (5-7) ، (5-8) ، (5-9+10) ، (5-11+12) ونضعها في الجدول (5-4) التالي، فنجد أن :

$$P(1) = \frac{N(1)}{n} = \frac{35}{35} = 1 \quad , \quad P(2) = \frac{N(2)}{n} = \frac{22}{35} \dots\dots\dots P(11) = \frac{N(11)}{n} = \frac{1}{35}$$

$$P(G_1/1 \in R(1)) = \frac{N_1(1)}{N(1)} = \frac{20}{35} \quad , P(G_2/x \in R(2)) = \frac{N_2(2)}{N(1)} = \frac{15}{35}$$

$$P[\ell(1)] = \frac{N[\ell(1)]}{n} = \frac{22}{35} \ , \ P[r(1)] = \frac{N[r(1)]}{n} = \frac{13}{35} \qquad \big(\text{غير موجودة في الجدول}\big)$$

$$P_{\ell 1} = \frac{P[\ell(1)]}{P(1)} = \frac{N[\ell(1)]}{N(1)} = \frac{22}{35} \ , \qquad P_{r1} = \frac{N[r(1)]}{N(1)} = \frac{13}{35} \dots\dots\dots\dots\dots\dots\dots$$

$$P_{\ell 2} = \frac{P[\ell(2)]}{P(2)} = \frac{N[\ell(2)]}{N(2)} = \frac{15}{22} \ , \qquad P_{r2} = \frac{N[r(2)]}{N(2)} = \frac{7}{22} \dots\dots\dots\dots\dots\dots\dots$$

جدول (5-4): احتمالات الانتماء والتوزيع :

t	قاعدة الفصل	$\ell(t)$	$r(t)$	$N(t)$	التوزيع على G_j		$P(t)$	$P(y_1/t)$	$P(y_2/t)$	P_ℓ	P_r
					$N_1(t)$	$N_2(t)$					
1	$x_2 \geq 2$	2	3	35	20	15	1	$\frac{20}{35}$	$\frac{15}{35}$	$\frac{22}{35}$	$\frac{13}{35}$
2	$x_1 < 2$	4	5	22	17	5	$\frac{22}{35}$	$\frac{17}{22}$	$\frac{5}{22}$	$\frac{15}{22}$	$\frac{7}{22}$
3	$X_1 \geq 1$	6	7	13	4	9	$\frac{13}{35}$	$\frac{4}{13}$	$\frac{9}{13}$	$\frac{9}{13}$	$\frac{4}{13}$
4	○	0	0	15	10	5	$\frac{15}{35}$	$\frac{10}{15}$	$\frac{5}{15}$	-	-
5	$X_2 \geq 3$	8	9	7	2	5	$\frac{7}{35}$	$\frac{2}{7}$	$\frac{5}{7}$	$\frac{5}{7}$	$\frac{2}{7}$
6	●	0	0	9	0	9	$\frac{9}{35}$	0	1	-	-
7	$x_2 < 1$	10	11	4	3	1	$\frac{4}{35}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{4}$
8	●	0	0	5	1	4	$\frac{5}{35}$	$\frac{1}{5}$	$\frac{4}{5}$	-	-
9	○	0	0	2	2	0	$\frac{2}{35}$	1	0	-	-
10	●	0	0	3	3	0	$\frac{3}{35}$	1	0	-	-
11	○	0	0	1	0	1	$\frac{1}{35}$	0	1	-	-

المصدر: Webb. A. R. (2002) Statistical Pattern Recigrition, P.231

ومن الجدول السابق نلاحظ أن العقد الخارجية هي (4) و(6) و(8) و(9) و(10) و(11)، ولقد تم تحديد انتماء كل منها إلى إحدى المجموعتين حسب الاحتمال الأكبر $P(y_j/t)$، في تلك العقدة . فمثلاً نجد أن العقدة (4) تقابل الاحتمالين $P(y_1/4) = \frac{10}{15}$ $P(y_2/4) = \frac{5}{15}$، فلذلك يتم تصنيف العقدة (4) إلى المجموعة الأولى G_1 المرموز لها الرمز ○، لأن الاحتمال المقابل لها $\left(\frac{10}{15}\right)$ هو أكبر الاحتمالين، وهكذا نقوم بتصنيف العقد الخارجية الأخرى .

5-5 معايير التفرع أو الانشطار :

لقد لاحظنا أن عملية التفرع تحتاج إلى قواعد ومعايير دقيقة تفرض على المتحولات X وعلى حدودها C في كل عقدة داخلية (t). فمثلاً يمكن أن نضع شرط التفرع كما يلي:

$$S_P = \left[X \in R^P \quad و \quad X_4 \leq 8{,}2\right] \qquad (14-5)$$

فنحصل على الفضاء الجزئي S_P من R^P الذي يحقق الشرط $X_4 \leq 8{,}2$.

وعندما يكون لدينا $X \in S_P$ فإن الجواب على الشرط $(X_4 \leq 8{,}2)$ يكون بـ (yes)، لذلك نصنف ذلك العنصر X على الفرع الأيسر باتجاه العقدة $\ell(t)$. أما إذا كان العكس فإننا نصنفه باتجاه العقدة اليمنى $R(t)$.

والسؤال الآن هو: بكم طريقة يمكننا أن نفرع البيانات X المتواجدة في المنطقة $R(t)$ المقابلة للعقدة t ؟

الجواب هو: بطرائق كثيرة ونحتاج إلى حساب معقدة وطويلة .

ويكفي أن نتذكر أن عدد طرائق التجزئة لأي مجتمع A يحتوي على k فئة خاصة، إلى مجموعتين غير خاليتين يساوي $(2^{k-1}-1)$ طريقة. عدا عن أنه يمكننا أن نضع في كل عقدة أحد المتحولات $X_1\ X_2\ \dots\ X_p$، وأن نضع عليه أي شرط ممكن وأن نختار له أي حد من مجاله C_i ...الخ، وهكذا يكون لدينا عدد كبير من الاختيارات في كل عقدة (t)، وذلك حسب قيم المتحول المختار X_i وحسب حدوده الممكنة C. وللخروج من هذا النفق تم وضع معايير للتفرع والانشطار عند كل عقدة t، ونقدم لها بمايلي:

لنفترض أنه لدينا في كل عقدة t عدة مخارج ممكنة هي: $G_1\ G_2\ \dots\ G_g$ وتشكل تابعاً نوعياً Y نرمز له بـ

$$Y: G_1\ G_2\ G_3\ \dots\ G_g \qquad (18-5)$$

وأنه لدينا P متحولاً تفسيرياً مؤثراً على Y نرمز لها بـ

$$X: X_1\ X_2\ X_3\ \dots\ X_p \qquad (19-5)$$

وإذا أخذنا أحد هذه المتحولات X وحددنا شرط التفرع عليه في العقدة t، ثم قمنا بحساب الاحتمالات المقابلة لتلك المجموعات من (5−8) و(5−9)، ورمزنا للتوزيع الاحتمالي المقابل لتلك المجموعات المتفرعة كمايلي:

$$\begin{aligned} &Y: G_1\ G_2\ G_3\ \dots\ G_g \\ &P(G_j/t): P_1\ P_2\ P_3\ \dots\ P_g \end{aligned} \qquad (20-5)$$

وهو يحقق الخاصتين التاليتين : $\sum_{j=1}^{g} P_j = 1$, $P_j \geq 0$

علماً بأن كل الاحتمالات P_j تحسب من العلاقة (5−8) التالية: $P_j = \frac{N_j(t)}{N(t)}$، وهنا نعرف على هذا التوزيع الاحتمالي المعايير التالية :

1− **معيار تابع الخطأ للتصنيف الخاطئ:** وهو مشتق من (5−13) ويسمى بتابع الشوائبية (عكس النقاوة) ويعرف في العقدة (t) بالعلاقة التالية :

$$Q_1(t) = 1 - \max_{j=1}[P_j] \qquad (21-5)$$

2−**مؤشر (جيني** $Gini$**)** والذي يعرف (في حالة التفرع الثنائي) وفي العقدة (t) بالعلاقة :

$$Q_2(t) = \sum_{j=1}^{g} P_j(1-P_j) = 1 - \sum_{j=1}^{g} P_j^2 \qquad (22-5)$$

3-**تابع القصور** $(entropy)$ **أو التشتت** $(deviance)$ ويعرف في العقدة (t) بالعلاقة:

$$Q_3(t) = -\sum_{j=1}^{g} P_j lg_2\left(P_j\right) \qquad (23-5)$$

حيث أن: lg_2 هو اللورغاريتم للأساس (2)، وهو يرتبط مع اللوغاريتم الطبيعي بالعلاقة :

$$lg_2 x = \frac{\ell n\, x}{\ell n\, 2}$$

حالة خاصة: إذا كان عدد المجموعات الممكنة عند العقدة t يساوي $g = 2$، فعندها يكون لدينا مخرجان فقط G_1 وG_2 ويقابلهما الاحتمالان P_1 و$(1 - P_1)$، وعندها فإن المقاييس الثلاثة السابقة تأخذ الشكل التالي :

$$Q_1(t) = 1 - \max[P_1\ , 1 - P_1] \qquad \text{تابع الخطأ} \qquad (24-5)$$

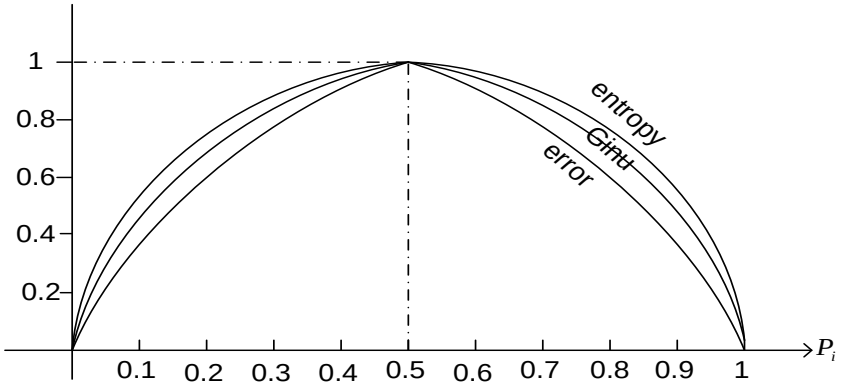
$$Q_2(t) = P_1(1 - P_1) + (1 - P_1) * P_1 = 2P_1(1 - P_1) \qquad (25-5)$$

$$Q_3(t) = -P_1 lg_2(P_1) - (1 - P_1) lg_2(1 - P_1) \qquad (26-5)$$

أما حالة الفشل فيعبر عنها بالعلاقة :

$$0 * lg_2 0 = 0 \qquad (27-5)$$

وأن هذه التوابع (للمجموعتين في التفرع الثنائي) ترسم المنحنيات والمستقيمات التالية :



الشكل (5-5): منحنيات المعايير السابقة في حالة مجموعتين G_1 و G_2

4- **التابع الاجمالي للشوائبية** : لتعريف التابع الاجمالي للشوائبية على الشجرة ككل، نفترض إننا نستخدم مقياس معين للشوائبية من المقاييس المذكورة سابقاً. وليكن $Q_i(t)$ على شجرة T، وهنا نعرف بعض المصطلحات ونرمز لها كما يلي:

- حجم الشجرة: وهو عدد العقد الخارجية فيها ونرمز له بـ $|T|$.
- المناطق المقابلة للعقد الخارجية في فضاء العينة R^P ونرمز لها بالرموز $R(1), R(2), \dots R(t) \dots R(|T|)$، وعددها يساوي $|T|$ منطقة منفصلة ومتكاملة .
- حجم عناصر العينة في كل منطقة $R(t)$ ونرمز له بالرمز N_t .
- قيمة تابع الشوائبية في العقدة t وليكن $Q(Rt)$.

وبناء على ذلك نعرف التابع الاجمالي للشوائبية في الشجرة ككل بواسطة العلاقة التالية:

$$Q = \sum_{t=1}^{|T|} N_t \, Q(R_t) \qquad (28-5)$$

وتحسب قيمته للشجرة ككل، للاستفادة منها في الحكم على جودة التفرع وعلى كفاية حجم أو عمق الشجرة . كما يمكننا استخدامه في عملية التوقف عن عمليات التفرع في كل مراحله وبدائله .

فكلما كانت قيمة Q صغيرة كانت جودة التفرع جيدة (علماً بأن أصغر قيمة له هي الصفر)، فعندما تبلغ قيمته أصغر من حد معين مثل (0 ,01) أو(0 ,05) نقرر التوقف عن متابعة التفرع ، ولكن هناك نقطة ضعف لاستخدام هذا التابع في عملية التوقف، وهي إنه عندما تكون الحدود صغيرة جداً أو معدومة القيمة تقودنا إلى تفريع متشعب جداً، وعندها تنشأ لدينا شجرة كبيرة جداً، ولكن بما أن الشجرة الكبيرة يمكن أن تشوه البيانات، وأن الشجرة الصغيرة يمكن أن تجسد فقط الهيكل العام للبيانات، فإنه يمكننا أن نضع ضابطاً آخر للتوقف، وهو أن نأمر بالتوقف عندما يبلغ عدد عناصر العينة في المنطقة R_t المقابلة للعقدة الخارجية t أقل من عدد محدد (5 مثلاً) أو عندما يساوي (1) حتماً .

5- تابع المكسب المعلوماتي *(Gain)*: وهناك معيار آخر (ولكنه معقد) لإجراء عمليات التفرع، يعتمد على حساب مقدار المكسب المعلوماتي الذي نحققه في كمية المعلومات اللازمة عند التفرع في كل عقدة t، وينطلق ذلك المعيار من حساب تابع القصور $(entropy)$ للتوزيع P_i بواسطة الفاصل X والحد C من العلاقة:

$$H(X\ ,C) = -\sum_{j=1}^{g} P_j * lg_2 P_j \qquad (29-5)$$

وهو يعبر عن كمية المعلومات اللازمة (الناقصة) لتوصيف حالة التفرع في العقدة t، ثم نقوم بحساب تابع القصور الشرطي في العقد المتفرعة عن t من العلاقة :

$$H_i(X/C) = -\sum_{j=1}^{g} P_j(X_i/C) * lg_2 P_j(X_i/C) \qquad (30-5)$$

حيث أن الاحتمالات الشرطية تحسب من العلاقة: $P_j(X/C) = \frac{N_j(t)}{N(t)}$

وبعدها نحسب التوقع الرياضي لـ $H_i(X/C)$ من العلاقة:

$$E[H_i(X/C)] = P_1(t_1')H_1 + P_2(t_2')H_2 + P_3(t_3')H_3 + \cdots \qquad (31-5)$$

وأخيراً يتم حساب المكسب المعلوماتي في قيمة تابع القصور، التي سيتم تحقيقها جراء التفريع عند العقدة t من العلاقة:

$$Gain\ (X) = \mathrm{H}(X,C) - E[\mathrm{H}_i(X/C) \qquad (32-5)$$

ثم نختار النتيجة التي تقابل أكبر قيمة للمكسب $(Gain)$، ونقرر اعتماد الفاصل $(X < C)$ للتفريع عند تلك العقدة .

مثال (5−4): لحساب المكسب المعلوماتي للتفرع عند عقدة الجذر (1) في المثال (5−3) المشروط بـ $(X_2 \geq 2)$، نجد أن معيار القصور يعطينا من الجدول (5−4) ما يلي:

$$H = -P_1 lg_2 P_1 - P_2 lg_2 P_2 = -\left[\frac{20}{35}\ln\frac{20}{35} + \frac{15}{35}\ln\frac{15}{35}\right] * \frac{1}{\ln 2}$$

$$H = \frac{0.319780 + 0.3631277}{0.69314718} = 0.985228 \ (bit)$$

ثم نقوم بحساب $H_1(X/C)$ و$H_2(X/C)$ في العقدتين (2) و(3) فنجد أن:

$$H_2 = -\frac{17}{22}\lg\frac{17}{22} - \frac{5}{22}\lg\frac{5}{22} = \frac{0.1992316 + 0.336728}{0.69314718} = 0.7732267$$

$$H_3 = -\frac{4}{13}\lg\frac{4}{13} - \frac{9}{13}\lg\frac{9}{13} = \frac{0.3626631 + 0.2545787}{0.69314718} = 0.890492$$

ثم نحسب التوقع الرياضي لهما فنجد أن:

$$E[\mathrm{H}_i(X/C)] = \frac{22}{35}\ H_1 + \frac{13}{35}\ H_2 = 0.4860282 + 0.3307542 = 0.816782 \ (bit)$$

وبذلك نجد أن مقدار المكسب المعلوماتي يساوي :

$$Gain\ (X/C) = H - E[\mathrm{H}(X/C)] = 0.985228 - 0.816782$$
$$Gain\ (X/C) = 0.168446 \quad (bit)$$

فإذا كانت قيمة هذا المكسب أكبر من جميع المكاسب الممكنة عن الشروط الممكنة للمتحولات X، نقرر اعتماد الشرط المستخدم (مثل $X_2 \geq 2$) للمتحول X_2 لتفرع البيانات والمشاهدات عند العقدة الأصلية (1). وهكذا نجد أن عملية تفريع العقدة (1) حسب$(X_2 \geq 2)$ إلى (2) و(3) جعلتنا نكسب (0.168446) بايتا (Bit) من كمية المعلوماتية، وهذا يزيد من قدرتنا على التصنيف الصحيح وتقلل من معدل التصنيف الخاطئ، وتجعلنا نحصل على عقدتين أكثر نقاوة من العقدة (1) .

ولكن حتى نحصل على نتيجة عامة تشمل جميع العقد الداخلية في الشجرة، يجب علينا أن نكرر مثل تلك الحسابات من أجل جميع شروط الفرز $(X \leq C)$، التي يمكن فرضها على المتحولات X وعند كل عقدة داخلية، وبما أن تلك الشروط قد تكون غير محدودة، لأنها تتعلق بعدد المتحولات X وبقيمتها العددية الممكنة وبالحدود المفروضة عليها C، فإن حجم الحسابات سيكون كبيراً جداً، ولا يمكن تنفيذه إلا بواسطة الحواسيب وباستخدام برنامج خاصة لذلك. ولكن يمكن تخفيض حجم هذه الحسابات باستخدام بعض الأساليب المفيدة. فمثلاً إذا كان X متحولاً مستمراً ضمن مجال معين فإننا نقوم بتبويب قيمه ضمن مجالات جزئية محدودة ونعتبره متحولاً مرتباً، ذا فئات محددة، وبذلك ينخفض حجم الحسابات السابقة كثيراً

ملاحظة 1: إن المعايير الثلاثة الأولى تتصف بالخواص التالية:

1−إن التوابع $Q(t)$ تأخذ أكبر قيمة لها وهي الواحد، وذلك عندما تكون الاحتمالات P_i متساوية. أي عندما: $P_1 = P_2 = P_3 = \cdots = \frac{1}{g}$، وهذه الحالة هي أسوأ الخيارات لأنها تجعل معدل التصنيف

الخاطئ أكبر ما يمكن، وهذا يقابل أعلى درجة للشوائبية وتساوي الواحد. فمثلاً عندما يكون لدينا Y مؤلفاً من مجموعتين فقط فإن $g = 2$، وعندما يكون الاحتمالان متساويان $P_1 = P_1 = 0.5$ فإن هذه التوابع تعطينا أكبر قيمة لتابع الشوائبية وتساوي الواحد، كما هو مبين على الشكل (5-5) .

2- إن التوابع $Q(t)$ تأخذ أصغر قيمة لها عندما يكون أحد الاحتمالات مساوياً للواحد، وتكون الاحتمالات الأخرى معدومة، أي عندما يأخذ التوزيع الاحتمالي أحد الأشكال التالية: (1,0,0,0,0) أو(0,1,0,0,0) أو(0,0,0,0,1)، وهذا يعني أن عناصر العينة تجمعت في مجموعة واحدة من مجموعات التابع Y، لأنها من صنف واحد، وبذلك تصبح درجة النقاوة 100% ويصبح تابع الشوائبية مساوياً للصفر $Q_n(t) = 0$.

3-إن التوابع$Q(t)$ تأخذ قيماً متناظرة مقابل الاحتمالات$(P_1 , P_2 , P_3,.., P_g)$ كما في الشكل(5-5) .

ملاحظة 2: عند استخدام أحد هذه المعايير في عمليات التفرع عند أية عقدة t، يتم اختيار المتحول X_k والشروط C_k (أو الفئات C_k)، التي تجعل قيمة المعيار المستخدم أصغر ما يمكن، وعادة يتم استخدام $Q_2(t)$ و$Q_3(t)$ في عمليات الإنشاء والتفرع عند العقد الداخلية . أما $Q_1(t)$ فيستخدم لإنشاء قاعدة للتوقف، لأنه يعبر عن جودة التصنيف عند العقدة (t)، كما يستخدم في عملية تقليم الشجرة حسبما سنرى لاحقاً .

5-6 التفريع بواسطة الاحتمالات السابقة والتكاليف : [انظر webb. A.R. P.238]

لقد تعرفنا في العلاقتين (5-7) (5-8) على الاحتمالات السابقة $P(t)$ و$P(j/t)$ ، والآن لنفترض أن الاحتمالات السابقة لكل مجموعة G_j يساوي $\pi(j)$ ويعرف بالعلاقة :

$$\pi(j) = \frac{N_j}{n} \quad , \quad P(j/t) = \frac{N_j(t)}{N_j} \qquad (33-5)$$

حيث أن: N_j هو عدد عناصر المجموعة G_j من أصل العينة ذات الحجم n. وهذه الاحتمالات هي عبارة عن نسبة المجموعة G_j في العينة المدروسة .

وإذا كان التوزيع الاحتمالي لمجموعة القرار لا يتناسب مع حدوث تلك المجموعة، فإن التقديرات الحصينة الكلية لاحتمالات أن تقع عناصر تلك العينة في داخل العقدة t يساوي $P(t)$، الذي يحسب من العلاقة (حسب نظرية الاحتمال الكلي) :

$$P(t) = \sum \pi(j) * \frac{N_j(t)}{N_j} \qquad (34-5)$$

حيث أن: $N_j(t)$ هو عدد عناصر المجموعة G_j الواقعة ضمن العقدة t، وعندما تكون G_j معطية مسبقاً فإن الاحتمالات اللاحقة تحسب من العلاقة:

$$P(t/j) = \frac{\pi(j) * \frac{N_j(t)}{N_j}}{\sum_{j=1}^{g} \pi(j) * \frac{N_j(t)}{N_j}} \qquad (35-5)$$

وإذا رمزنا لنسبة عناصر العينة التي من المجموعة G_j وصنفت خطأ في المجموعة G_i بالرمز $q(i/j)$، وتجاهلنا تكاليف تلك الأخطاء (أو اعتبرناها متساوية) فإن معدل التصنيف الخاطئ لكامل الشجرة يحسب من العلاقة :

$$R(T) = \sum_{i,j}^{n} q(i/j) * \pi(j) \qquad (36-5)$$

وبعد حساب قيم $R(T)$ نختار أصغر القيم الممكنة لها لاختيار التفريع الأفضل .

وإذا افترضنا أن تكاليف التصنيف الخاطئ لأي عنصر من المجموعة G_j في المجموعة G_i تساوي C_{ij}، فعندها نجد أن تابع تكاليف التصنيف الخاطئ على كامل الشجرة يساوي :

$$C(T) = \sum_{i\ j}^{n} C_{ij} * q(i/j) * \pi(J) \qquad (37-5)$$

ثم نقوم بحساب قيم هذا التابع، ونختار الحالة التي يأخذ فيها أصغر قيمة ممكنة لاختيار التفريع الأقل تكلفة .

5-7 التفريع بواسطة الانحدار التجميعي (MARS): (انظر Friadman P.1991 أو webb. P.242)

وهي مأخوذة من الكلمات الإنكليزية (Multivoriate Adaptive Regression Splines) لنفترض أنه لدينا بيانات مؤلفة من n قياساً لـ p متحولاً هي $(X_1, X_2, X_3, \dots, X_p)$ ويقابلها n قياساً لتابع الاستجابة Y وتشكل مصفوفة من المرتبة $(p, n+1)$، كما نفترض أنه يمكننا توليد أو تمثيل هذه البيانات بواسطة علاقة انحدار من الشكل :

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{pi}) + \varepsilon_i \qquad (38-5)$$

حيث أن ε_i هو حد الخطأ المرتكب (البواقي)، والمطلوب منا أن نجد تقديراً معيناً لـ f مثل $\tilde{f}$ بحيث يكون الخطأ ε_i أقل ما يمكن .

إن نموذج التجزئة المتكررة المتتالية الثنائية يعرف بواسطة العلاقة التالية :

$$\tilde{f}(X) = \sum_{t=1}^{|T|} a_t B_t(X) \qquad (39-5)$$

حيث أن التابع الأساسي $B_t(X)$ يساوي :

$$B_t(X) = I[x \in R(t)] \qquad (40-5)$$

حيث أن I هو تابع ثنائي يأخذ القيمة (1) إذا كان الشرط $x \in R(t)$ صحيحاً، ويأخذ القيمة (0) إذا كان الشرط غير صحيح .

وحيث أن $R(t)$ هي المناطق المفصولة، المقابلة للعقد الخارجية وعددها $|T|$ ... 3 2 1 :i :t وهي تحقق العلاقتين: $R_t \cup R_j = \phi$ من أجل $t \neq j$. وأن $\cup R_j = \Omega$.

أما مجموعة الرموز a_t (حيث لـ $t = 1,2$) فهي الأمثال العددية للتابع (5−39) والتي تحسب قيمها بطريقة المربعات الصغرى أو الإمكانية العظمى وذلك لتمثيل تلك البيانات .

ولكن في مسائل التصنيف فإنه يمكن إعادة صياغة تابع الانحدار في كل مجموعة (باستخدام المتحولات الثنائية التي تأخذ إحدى القيمتين (1) إذا كانت $x_i \in G_j$ وتأخذ الصفر إذا حدث غير ذلك) إلى عدة توابع $\tilde{f}_j$ حسب القواعد المتبقية في التمييز . وهكذا يمكننا من إنتاج التابع الأساسي بواسطة خوارزمية (*Friedman 1991*) وتقديمه كمنتج مؤلف من جداء عدة توابع خطية .

ولتوضيح ذلك نأخذ التجزئة الناتجة عن الشجرة المعطية في المثال (5−2)، فنجد أن التجزئة الأولى قامت على المتحول X_2 وقسمت المستوى حسب الشرط $(X_2 \geq b_2)$ إلى منطقتين وأعطتنا التابعين التاليين :

$$H[(X_2 - b_2)] \quad , \quad H[-(X_2 - b_2)] \qquad (41-5)$$

حيث أن التابع $H(X)$ يساوي :

$$H(X) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{غير ذلك} \end{cases} \qquad (42-5)$$

حيث أن: $x = X_2 - b_2$

ثم تمت تجزئة المنطقة المقابلة $(X_2 < b_2)$ مرة ثانية استناداً إلى المتحول X_1 والشرط $(X_1 < a_1)$، وهذا يعطينا تابعين آخرين أساسيين يشكلان مع التابع الثاني السابق تابعين جديدين (بواسطة تقاطعها) ويمكن كتابتهما على الشكل التالي :

$$H[-(X_2 - b_2)] * H[-(X_1 - a_1)] \quad \text{و} \quad H[-(X_2 - b_2)] * H[+(X_1 - a_1)] \qquad (43-5)$$

وهكذا يمكننا أن نحصل على عدة توابع أساسية نهائية لكامل الشجرة تتألف من الجداءات التالية:

$$\begin{aligned} &H[-(X_2 - b_2)] * H[-(X_1 - a_1)] * H[+(X_2 - b_1)] \\ &H[-(X_2 - b_2)] * H[-(X_1 - a_1)] * H[-(X_2 - b_1)] \\ &H[-(X_2 - b_2)] * H[+(X_1 - a_1)] \\ &H[+(X_2 - b_2)] * H[-(X_1 - a_2)] \\ &H[+(X_2 - b_2)] * H[+(X_1 - a_2)] * H[+(X_2 - b_3)] \\ &H[+(X_2 - b_2)] * H[+(X_1 - a_2)] * H[-(X_2 - b_3)] \end{aligned} \qquad (44-5)$$

وهنا نلاحظ أن كل تابع من هذه التوابع هو عبارة جداء عدة توابع H، وبصورة عامة فإن التوابع الأساسية الناتجة عن خوارزمية التجزئة المتتالية يكون لها الشكل التالي :

$$B_t(X) = \prod_{k=1}^{kt} H\left[S_{kt}\left(X_{v(kt)} - C_{kt}\right)\right. \qquad (45-5)$$

حيث أن S_{kt} هو تابع الإشارة ويأخذ إحدى القيمتين (± 1)

أما kt فهو عدد التفريعات التي تؤدي بنا إلى $B_m(X)$

وأن $X_{v(kt)}$ فهو متحول التفريع، والرمز C_{kt} هو قيمة حد الفصل المفروض على المتحول X .

إن أسلوب MARS هو توليد لأسلوب التجزئة المتتالية حسب الطريقة التالية .

- مسألة الاستمرارية في MARS :

إن نموذج التجزئة المتتالية MARS ينقطع في منطقة الحدود . وهذا يكون بسبب تابع الخطوة H، ولأن أسلوب MARS يستبدل توابع الخطوة بواسطة توابع التفريع . وإن الجانبين يشكلان توابع أسية أساسية للفواصل من المرتبة q كما يلي:

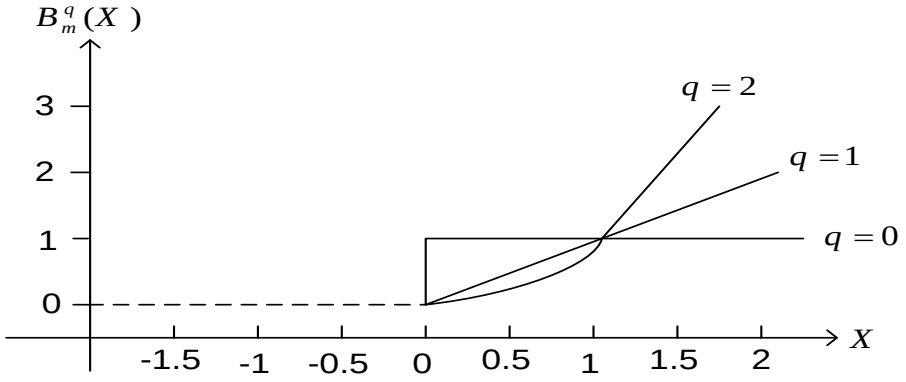
$$b_q^{\pm}(X-t) = [\pm(X-t)]_+^q \qquad (46-5)$$

حيث أن الرمز $[\ \ ]_+$ يرمز إلى الجزء الموجب للقطعة المعتبرة . أما التابع الأساسي $b_q^+(X)$ فهو موضح في الشكل (5-6) . وبذلك يكون تابع الخطوة H هو حالة خاصة للقيمة $q = 0$. كما يمكن استخدام خوارزمية MARS عندما $q = 1$، وهذا يؤدي إلى استمرارية تابع التقريب ولكنه يتقاطع مع مشتقاته الأولى .

وعندها فإن التابع الأساسي يأخذ الصيغة التالية :

$$B_t^q(X) = \prod \left[S_{kt}\left(X_{v(kt)} - C_{kt}\right)\right]^q \qquad (47-5)$$

حيث أن C_{kt} تشير إلى زمرة الحدود الفاصلة .



الشكل (5-6) توابع التفريع من أجل $q = 0$ $q = 1$ $q = 2$

5-8 التقليم (Pruning): [بتصرف webb. A.R, P.233] .

إن التقليم يطبق على أشجار التصنيف الجاهزة، للتخلص من الفروع الزائدة، ولمعالجة هذه المسألة نفترض أنه يوجد عند كل عقدة t من الشجرة المعطية T، عدد حقيقي $q(t)$، يمثل احتمال التصنيف الخاطئ في العقدة t، فإذا كانت t عقدة خارجية (أي أن $t \in |T|$) وكان $M(t)$ عدد عناصر العينة، الذين تم تصنيفهم خطأ في تلك العقدة الخارجية t، فإن $q(t)$ يعبر عن نسبة عدد عناصر العينة من المنطقة $R(t)$، الذين لا ينتمون إلى الفئة المقابلة لتلك العقدة الخارجية t. وإن $q(t)$ يحسب من العلاقة التالية:

$$q(t) = \frac{M(t)}{n} : \quad t \in |T| \qquad \text{للعقد الخارجية :} \qquad (48-5)$$

حيث أن: $|T|$ هو عدد العقد الخارجية في الشجرة المدروسة، وبذلك نجد أن معدل التصنيف الخاطئ لكامل الشجرة T يساوي مجموع معدلات الخطأ في العقد الخارجية فيها، ولنرمز له بـ $Q(T)$ ونحسبه من العلاقة :

$$Q(T) = \sum_{t\in|T|} q(t) \qquad (47-5)$$

وهو يعبر عن تابع الشوائبية (عدم النقاء) في الشجرة الكلية T.

ولنفترض الآن أن هناك تكاليف أخرى قد تفرض على العقد الخارجية للشجرة T. (مثل الضرائب أو الغرامات أو تخفيض الأسعار أو التلف)، وإن قيمة كل منها تساوي $\propto$ (قيمة موحدة لجميع العقد الخارجية)، فعندها نجد أن معدل خطأ التصنيف المركب لكل عقدة خارجية t قد أصبح يساوي:

$$q_{\propto}(t) = q(t)+\propto \qquad (48-5)$$

وبذلك يصبح معدل خطأ التصنيف المركب لكامل الشجرة T مساوياً لـ $Q_{\propto}(t)$ ويحسب من العلاقة :

$$Q_{\propto}(t) = \sum_{t\in|T|} q_{\propto}(t) = \sum_{t\in|T|} [q(t)+\propto]$$

وبناء على (5-47) نجد أن:

$$Q_{\propto}(t) = Q(T)+\propto |T| \qquad (49-5)$$

وبما أن عملية التقليم تهدف إلى حذف بعض فروع الشجرة T والعقد الخارجية المتعلقة بها، وإن هذه الفروع تشكل شجيرات فرعية، وتكون أصولها في أحد العقد الداخلية t، لذلك يجب إعادة حساب $q(t)$ و $Q(T)$ و $q_{\propto}(t)$ و $Q_{\propto}(t)$ بعد كل عملية تقليم، وذلك من أجل التمهيد لعملية التقليم التالية .

والآن لنرمز للشجرة الفرعية التي أصلها في العقدة الداخلية t من الشجرة T بالرمز T_t. ولعدد العقد الخارجية المتعلقة بها بـ $|T_t|$، ولمعدل خطأ التصنيف المركب فيها بالرمز $Q_{\propto}(T_t)$. وبطريقة مشابهة للمعالجة السابقة نجد أنه قياساً على (5-49) أن:

$$Q_{\propto}(T_t) = Q(T_t)+\propto |T_t| \qquad (50-5)$$

ومن جهة أخرى نجد أن العقدة t بعد التقليم ستصبح عقدة خارجية في الشجرة الجديدة، وبالتالي فإن معدل خطأ التصنيف المركب فيها يصبح مساوياً لـ

$$q_{\propto}(t) = q(t)+\propto \qquad (51-5)$$

وهو عبارة عن معدل خطأ التصنيف المركب للشجرة المتفرعة من العقدة t، ولذلك فإنه بعد عملية التقليم يجب أن يتساوى المعدلان $Q_{\propto}(T_t)$ و $q_{\propto}(t)$، لأنهما يعبران عن معدل خطأ التصنيف المركب لنفس الشجرة الفرعية T_t. لذلك نضع بينهما المعادلة التالية:

$$Q_{\propto}(T_t) = q_{\propto}(t)$$
$$Q(T_t)+\propto |T_t| = q(t)+\propto$$

ومنها نجد أن $\propto$ تساوي:

$$\propto = \frac{q(t) - Q(T_t)}{|T_t| - 1} \qquad (52-5)$$

فإذا كانت قيمة $\propto$ صغيرة بالنسبة لشجيرة محددة T_t، فهذا يعني أن الفرق بين $[q(t) - Q(T_t)]$ يكون صغيراً، وبالعكس. وعندها يكون لدينا شجرة فرعية كبيرة ولها عدد كبير $|T_t|$ من العقد الخارجية التي ستجعل المقدار $Q(T_t)$، قريباً من $q(t)$، وهذا ما يجعل عملية التقليم مجدية .

أما عندما تكون قيمة $\propto$ كبيرة نسبياً فإنها تقابل شجيرة فرعية صغيرة T_t من T، ويكون لها عدد قليل من العقد الخارجية $|T_t|$، وبالتالي يكون الفرق $[q(t) - Q(T_t)]$ كبيراً نسبياً، وعندها تكون عملية التقليم ليست ذات أهمية .

لذلك اقترح (بريمان Breiman 1984) تعريف تابع جديد $g(t)$ مشابهاً للعلاقة (5-52)، لاستخدامه في تحديد العقدة t، التي تصلح لأن تكون أصلاً لشجرة فرعية T_t، وهي العقدة التي تقابل أصغر قيمة لذلك التابع $g(t)$. وإن $g(t)$ يساوي:

$$g(t) = \frac{q(t) - Q(T_t)}{|T_t| - 1} \qquad (53-5)$$

ولكنه قدم تعريفاً آخر لحساب المعدل $q(t)$ وهو:

$$q(t) = P(t)\left[1 - \max P(G_j/t\right] \qquad (54-5)$$

ونلخص خوارزمية (Breiman 1984) للتقليم بما يلي:

أولاً نتأكد من أن شجرة التصنيف جاهزة للتقليم ثم نقوم بما يلي:

1-نقوم بحساب معدلات خطأ التصنيف العادي $q(t)$ لجميع العقد الداخلية والخارجية في الشجرة من العلاقة (5-54)، ونسجله على يسار العقد الداخلية ونضعه بأسفل العقد الخارجية (كما في الشكل (5-7)) .

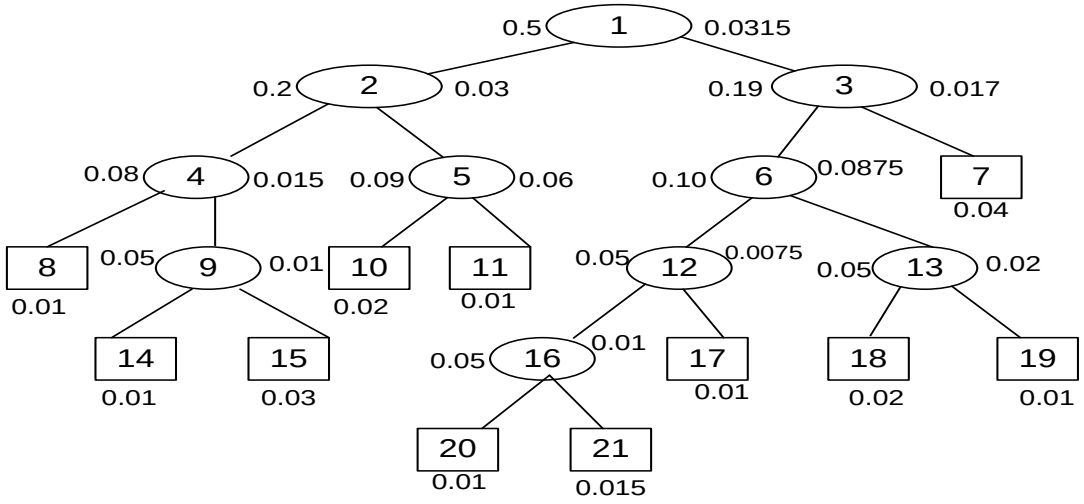
2-نقوم بحساب قيم التابع $g(t)$ لجميع العقد الداخلية في الشجرة من العلاقة (5-53) ونسجلها على يمين العقد الداخلية (كما في الشكل (5-7)) .

3-نبحث عن أصغر قيمة للتابع $g(t)$ ومنها نحدد العقدة المقابلة t التي ستكون أصلاً للشجيرة الفرعية T_t، التي يجب تقليمها، ثم نحدد العقد المتفرعة عنها استعداداً لحذفها (يمكن أن تصادف أكثر من عقدة t وأكثر شجيرة فرعية T_t) .

4-نحذف العقد المتفرعة عن العقدة t المحددة في (3) ونعيد رسم الشجرة المتبقية مع الحفاظ على أرقام العقد السابقة وعلى قيم المعدلات $q(t)$ المسجلة على يسار كل عقدة داخلية لأنها لا تتغير .

5-نعود إلى الخطوة (2) ونكرر إعادة حساب قيم التابع $g(t)$ لجميع العقد الداخلية في الشجرة المتبقية، ونكرر هذه الحلقات حتى نتوصل إلى الشجيرة الفرعية المؤلفة من العقدة الأصلية (الجذر) فقط .

مثال (5-5) : الآن لنأخذ شجرة التصنيف التالية :



الشكل (5-7) شجرة تصنيف افتراضية

ومن الشكل (5-7) نلاحظ أن كل عقدة خارجية قد ميزت برقم واحد هو قيمة الاحتمال $q(t)$ المحسوبة من (5-54)، وهو عبارة عن احتمال التصنيف الخاطئ، الذي يعبر عن نسبة مساهمة تلك العقدة t في معدل الخطأ الاجمالي، ولكن كل عقدة داخلية قد ميزت برقمين كتبا على يسارها ويمينها وهما:

إن الرقم المكتوب على يسار العقدة t هو قيمة المقدار $q(t)$، وهو يعير عن نسبة مساهمة تلك العقدة في معدل الخطأ الاجمالي، فيما إذا أصبحت تلك العقدة عقدة خارجية (ورقة). وهنا نلاحظ أن قيم $q(t)$ تتناقص كلما تفرعت الشجرة إلى فروع أو أوراق متجانسة أو صافية وتصبح قيمه صغيرة جداً في العقد الخارجية .

أما الرقم المكتوب على يمين العقدة t، فهو قيمة التابع $g(t)$ المحسوب من العلاقة (5-53)، فمثلاً نجد أن قيمة $g(t)$ عند العقدة $t = 2$ تحسب على العقد الخارجية: (8) و(10) و(11) و(14) و(15) من العلاقة (5-53) فنجد أن :

$$g(2) = \frac{q(2) - Q(T_2)}{|T_2| - 1} = \frac{0.2 - (0.01 + 0.01 + 0.03 + 0.02 + 0.01)}{5 - 1} = \frac{0.12}{4} = 0.03$$

أما في العقدة $t=3$ فإن $g(3)$ يحسب على العقد الخارجية: (7) و(17) و(18) و(19) و(20) و(21) فنجد من العلاقة (5-53) أن :

$$g(3) = \frac{q(3) - Q(T_3)}{|T_3| - 1} = \frac{0.19 - (0.04 + 0.01 + 0.02 + 0.01 + 0.01 + 0.015)}{6 - 1} = \frac{0.085}{5} = 0.017$$

وكذلك نجد أن قيمة $g(t)$ في العقدة $(t = 1)$ تحسب على جميع العقد الخارجية في الشجرة T وهي تساوي:

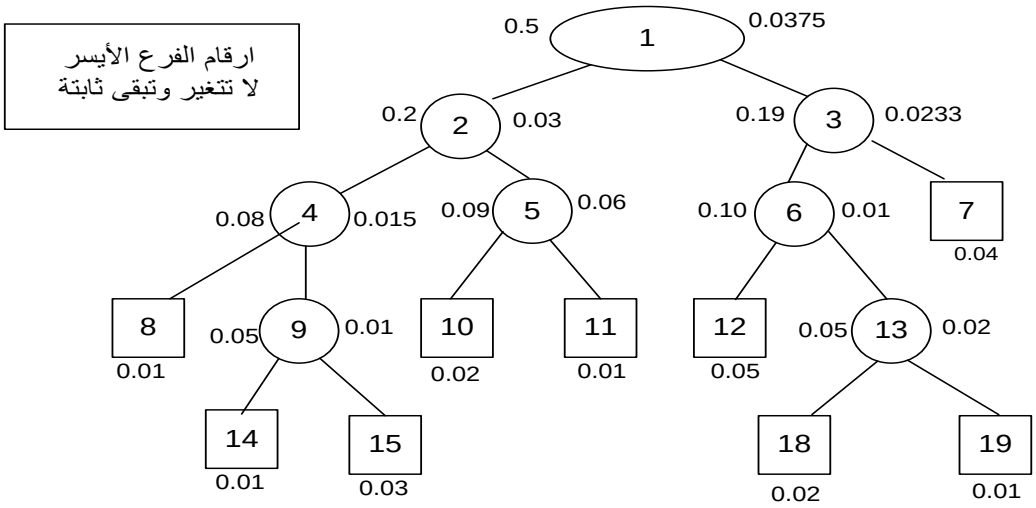
$$g(1) = \frac{0.5 - (0.01 + 0.01 + 0.03 + 0.02 + 0.01 + 0.04 + 0.01 + 0.015 + 0.01 + 0.02 + 0.01)}{11 - 1} =$$

$$= \frac{0.5 - 0.185}{10} = 0.0315$$

ولإجراء عملية التقليم نتبع الخطوات التالية :

1- ندرس جميع قيم التابع $g(t)$ المكتوبة على يمين العقد الداخلية ثم نحدد أصغرها، فتكون العقدة الداخلية t المقابلة لأصغر قيمة لـ $g(t)$ هي العقدة المرشحة لتكون عقدة خارجية (ورقة)، لذلك نعتبر تلك العقدة عقدة خارجية ونحذف جميع الفروع والعقد الصادرة عنها، ونرمز للشجرة الناتجة عن هذه الخطوة الأولى بـ T^1 . وفي مثالنا هذا نجد أن أصغر قيمة لـ $g(t)$ هي 0.0075 المقابلة للعقدة (12)، لذلك نعتبر هذه العقدة عقدة خارجية ونحذف جميع الفروع والعقد الصادرة عنها، ونضع قيمة $q(t)$ تحتها . علماً بأن $q(12) = 0.05$

2- نعود ونقوم بحساب قيم التابع $g(t)$ لجميع العقد الداخلية السابقة لتلك العقدة (12) (أسلافها)، بما في ذلك عقدة الجذر الأصلي (1)، فنحصل على الشجرة المقابلة لهذه الخطوة وهي الشجرة T^1 التالية :



الشكل (5-8) نتيجة التقليم الأول T^1

وهنا نشير إلى أن قيم $g(6)$ و $g(3)$ و $g(1)$ قد تم حسابهما كما يلي :

$$g(6) = \frac{0.10 - (0.05 + 0.02 + 0.01)}{3 - 1} = 0.01$$

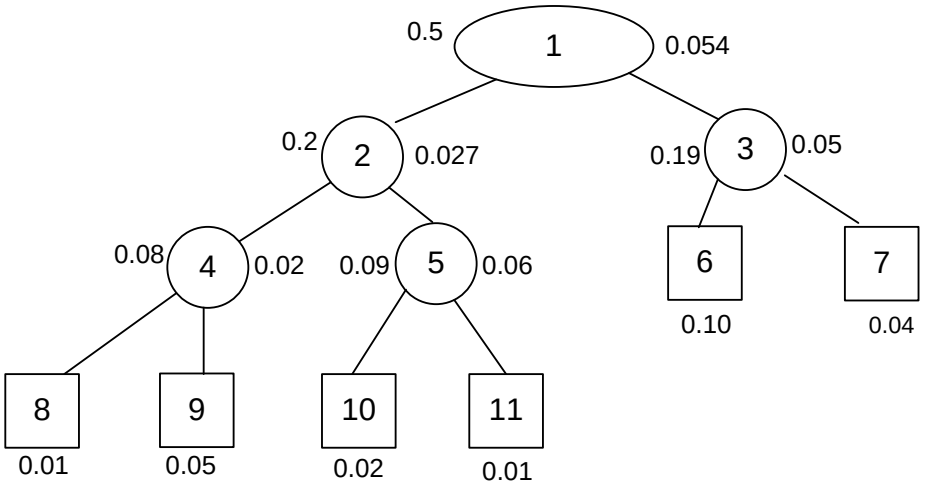
$$g(3) = \frac{0.19 - (0.05 + 0.02 + 0.01 + 0.04)}{4 - 1} = 0.02333$$

$$g(1) = \frac{0.5 - (0.05 + 0.02 + 0.01 + 0.04 + 0.01 + 0.01 + 0.03 + 0.02 + 0.01)}{9 - 1} = 0.0375$$

أما قيم $g(t)$ على العقد الداخلية التي في الفرع الأيسر فتبقى ثابتة كما كانت عليه في الشكل $(5-7)$ السابق .

3- نعود الآن إلى الخطوة (1) وندرس من جديد قيم التابع $g(t)$ عند جميع العقد الداخلية المتبقية، ونحدد أصغر تلك القيم، ونعتبر العقدة الداخلية (أو العقد) المقابلة لها مرشحة لأن تكون عقدة خارجية، ثم يتم حذف العقد المتفرعة عنها. وفي هذه الخطوة نجد أن أصغر قيم $g(t)$ هي $g(6) = 0.01$ وكذلك $g(9) = 0.01$ (من الفرع الأيسر) .

لذلك نعتبر هاتين العقدتين عقدتين خارجيتين (ورقتين) ونحذف العقد التي تتفرع عنهما، فنحصل على الشجرة المقابلة لهذه الخطوة، والتي سنرمز لها بـ T^2 التالية :



الشكل (5–9) نتيجة التقليم الثاني T^2

ثم نقوم بحساب قيم $g(t)$ المقابلة للعقد السابقة للعقدتين (6) و(7) فنجد مثلاً أن:

$$g(3) = \frac{0.19 - (0.10 + 0.04)}{2 - 1} = 0.05$$

$$g(1) = \frac{0.5 - (0.10 + 0.04 + 0.01 + 0.05 + 0.02 + 0.01)}{6 - 1} = 0.054$$

وكذلك حسبنا قيم $g(t)$ المقابلة للعقد الأخرى على الفرع الأيسر ووضعناها على الشكل (5–9) فحصلنا على الشجرة T^2 .

4- نعود مرة أخرى إلى الخطوة (1) ونكرر تعليماتها، فنجد أن أصغر قيمة لـ $g(t)$ هي (0,02) التي تقابل العقدة الداخلية (4)، لذلك نجعل العقدة (4) عقدة خارجية، ونحذف العقدتين المتفرعتين عنها (8) و(9) من الشجرة T، فنحصل على الشجرة المقلمة T^3. وهكذا نتابع ونعود ونكرر تعليمات الخطوة (1)، ونكرر الحسابات، ثم نبحث عن العقدة الداخلية (العقد) المقابلة لأصغر قيمة لـ$g(t)$ فنجد أنها هي المقابلة للعقدة (2)، لذلك نحذف العقد (4) و(5) و(10) و(11) فنحصل على الشجرة المقلمة T^4 التي لا تحتوي على العقد (4) و(5) و(10) و(11) . ثم نكرر ذلك فنحصل على الشجرة المقلمة T^5، التي لا تحتوي على العقدتين (6) و(7) . وأخيراً نحصل على الشجرة المقلمة T^6 التي تتألف من العقدتين (2) و(3) فقط . وإذا حذفنا العقدتين (2) و(3) نحصل على الشجرة المقلمة الأخيرة، التي تتألف من عقدة واحدة فقط هي عقدة الأصل .

وأخيراً نلخص نتيجة عمليات التقليم المتتالية للشجرة السابقة (5–7) في الجدول التالي :

جدول (5–5) نتائج التقليم المتتالي من T^1 حتى $T^{\bar{x}}$

رقم الشجرة المقلمة	العقد التي يجب حذفها من T^{k-1}	احتمال الخطأ الاجمالي $Q(T^{k-1})$	عدد العقد الخارجية $\|T\|$	رقم العقدة المقابلة للأصغر $\propto$	أصغر قيمة لـ $\propto$ أو للتابع $g(t)$	رقم الخطوة k
T^0	قيد الدراسة	0.185	11	البداية	0	0
T^1	10+17+20+2	0.20	9	12	0.0075	1
T^2	12+13+14+ 15+18+19	0.22	6	6+9	0.01	2
T^3	8+9	0.25	5	4	0.02	3
T^4	4+5+10+11	0.34	3	2	0.045	4
T^5	6+7	0.39	2	3	0.05	5
T^6	2+3	0.50	1	1	0.11	6

ملاحظة: نلاحظ أن احتمالات الخطأ الاجمالي المبين في العمود الخامس من الجدول (5–5) تتزايد كلما ازدادت خطوات التعليم. وهذا يعني أن التقليم الشديد يزيد من احتمال الخطأ الاجمالي لذلك فإن عملية التقليم تطبق على الأشجار الكبيرة، وتتوقف عن حد معين لاحتمال الخطأ الاجمالي .

5–9 طرائق اختبار جودة تصميم شجرة التصنيف [Webb. 2002 P.236]:

بعد إنجاز عمليتي التفريع والتقليم لشجرة التصنيف وتقدير معدل الخطأ فيهما . سنستعرض بعض طرائق اختبار جودة التفريع والتقليم وأهمها:

1- طريقة استقلال التجارب:

لنفترض أنه لدينا مجموعة بيانات ميدانية L_t ومجموعة من الاختبارات L_s لمصداقية بيانات العينة، وهذه المجموعة تتجمع بواسطة مجموعة من القرارات ضمن مجموعتين تقريباً . وذلك وفق الخطوات التالية :

أ- نستخدم المجموعة L_t لتشكيل الشجرة T بواسطة تفريع جميع العقد الممكنة، حتى تصبح جميع العقد الخارجية نقية وصافية، أي حتى تكون جميع عناصر العينة في كل عقدة خارجية تنتمي إلى مجموعة واحدة (فئة واحدة)، ولكن هذا يمكن أن لا يتحقق بسبب تداخل التوزيعات الاحتمالية. لذلك يمكن اتباع بديل آخر لإيقاف التفريع وذلك عندما يصبح عدد العناصر في كل عقدة خارجية أقل من عدد محدد مسبقاً، أو عندما يصبح عدد العناصر في العقدة اليسرى أو اليمنى مساوياً للصفر أو الواحد .

ب- نستخدم خوارزمية التقليم لتوليد أعشاش (خصائل) من الأشجار الفرعية نرمز لها بـ T^k، وذلك اعتماداً على مجموعة الاختبارات L_s .

ج- نختار الشجرة الفرعية T^k التي تكون قيمة تابع الشوائبية $Q(T^k)$ فيها أصغر ما يمكن .

2- طريقة المصداقية الاجمالية (Cors - Validation) :

إن طريقة المصداقية الاجمالية تقتضي أن نقوم بتقسيم المجموعة L_t إلى v مجموعة أو منطقة منفصلة هي:

$$L_1 , L_2 , L_3 , \dots , L_v \qquad (55-5)$$

وبحيث تكون أحجامها متساوية تقريباً (حجم كل منها يساوي $\frac{n}{v}$) .

ثم نفترض أن الفرق بين المجموعة الكلية L وأية مجموعة جزئية L_v يساوي :

$$L^v = L - L_v \qquad v: 1\ 2\ 3\ \ldots\ldots v \qquad (56-5)$$

وسنرمز بـ $T(\ltimes)$ للشجرة الفرعية المقلمة، التي تكون كل العقد فيها لديها قيمة للتابع $g(t)$ تحقق الشرط التالي $g(t) < \ltimes$، حيث $\ltimes$ هو أصغر قيمة التابع $g(t)$ في الشجرة . وعندها تكون الشجرة $T(\ltimes)$ مساوية للشجرة T^k (الشجرة الفرعية المقلمة في المرحلة k) . وحيث أن k يتم اختياره أو تحديده بحيث يحقق العلاقة التالية:

$$\ltimes_k \leq \ltimes \leq \ltimes_{k+1} \qquad : (\ltimes_{k+1} \rightarrow \infty) \qquad (56-5)$$

وعندها نجد أن خطوات اختبار المصداقية الاجمالية تكون كما يلي:

1- نستخدم المجموعة L_t لتوليد الشجرة T بواسطة تفريع كل العقد كما ورد في طريقة استقلال التجارب السابق .

2- نستخدم خوارزمية التقليم لتوليد أعشاش (خصائل) من الأشجار المقلمة نرمز لها على الترتيب كما يلي:

$$T = T^0 \geq T^1 \geq T^2 \geq T^3 \geq \cdots \geq T^k = root(k) \qquad (57-5)$$

3- نستخدم المجموعة الجزئية L_v لتوليد الشجرة الفرعية T_v، ونحدد المجموعة التي تصنف في العقدة الخارجية المناسبة، وذلك من أجل : $v: 1\ 2\ 3\ \ldots\ldots v$.

4- نستخدم خوارزمية التقليم لتوليد أعشاش من الأشجار الفرعية المقلمة T_v .

5- نحسب قيمة مؤشر المصداقية الاجمالية $Q^{cv}(T^k)$ (تقدير المصداقية الاجمالية لمعدل التصنيف الخاطئ) من العلاقة :

$$Q^{cv}(T^k) = \frac{1}{V}\sum_{v=1}^{V} Q_v\left[T_v\left(\sqrt{\ltimes_k * \ltimes_{k+1}}\right)\right] \qquad (58-5)$$

حيث أن: Q_v هو تقدير معدل التصنيف الخاطئ في المنطقة L_v، وذلك للشجرة الفرعية $T_v\left(\sqrt{\ltimes_k * \ltimes_{k+1}}\right)$.

6- نختار الشجرة الفرعية المقلمة T^* الصغرى التي يكون فيها :

$$Q^{cv}(T^*) = \min[Q^{cv}(T^k) \qquad (59-5)$$

7- نقدر معدل التصنيف الخاطئ من العلاقة :

$$\tilde{Q}(T^*) = Q^{cv}(T^*) \qquad (60-5)$$

إن الإجراءات السابقة تستخدم في عدة تطبيقات لشجرة التصنيف . وذلك خلال عمليات الإنشاء والتقليم، ويمكن دمجهما في تقنية واحدة للحصول على الحجم الصحيح للشجرة الناتجة عن المتحولات المستمرة أو المنقطعة أو المرتبة أو الأسمية أو المتضمنة مجموعات بيانات مختلفة من أنواع هذه المتحولات .