اكتشاف قواعد القرار المحددة وغير المحددة من قواعد المعطيات

د. عبد الله حسن العمر*

الملخص

يعتبر اكتشاف المعرفة من قواعد المعطيات منطقة بحثية نشطة واعدة تبشر بربح وافر في العديد من تطبيقات الأعمال والتطبيقات العلمية. إن التحدي الرئيسي لاكتشاف المعرفة من قواعد المعطيات هو المعالجة الآلية لكميات المعطيات الخام والكبيرة، وتحديد الأنماط الأكثر أهميةً والتي لها معنى، وتقديم هذه المعرفة في شكل مناسب لتحقيق هدف المستخدم. تواجه نظم اكتشاف المعرفة مشكلات تحدٍ من قواعد معطيات العالم الحقيقي التي تتجه لأن تكون كبيرة جداً، ومكررة، ومفعمة بالضجيج، وديناميكية. ولقد تم تناول كل مشكلة من هذه المشكلات إلى حد معين ضمن تعلم الآلة. معالجة بعض هذه المشكلات أثناء إنتاج معرفة مفيدة بشكل فعال ومؤثر هي المركز الرئيسي لهذه المقالة. ومن أجل معالجة بعض هذه المشكلات كان هدفنا الرئيسي في هذه المقالة هو كيفية استغلال نظرية المجموعات التقريبية في اكتشاف قواعد القرار المحددة وغير المحددة من قواعد المعطيات.

الكلمات المفتاحية: اكتشاف المعرفة، نظرية المجموعات التقريبية، قواعد القرار المحددة، قواعد القرار غير المحددة، قاعدة المعطيات.

* مدرس – قسم العلوم الأساسية– كلية الهندسة المعلوماتية– جامعة دمشق.

Discovering Deterministic and Nondeterministic Decision Rules in Databases

Dr. Abdullah Hassan Alomar*

Abstract

Knowledge Discovery in Databases (KDD) is an active research area with the promise for a high profit in many business and scientific applications. The grand challenge of knowledge discovery in databases is to automatically process large quantities of raw data, identify the most significant and meaningful patterns, and present this knowledge in an appropriate form for achieving the user's goal. Knowledge discovery systems face challenging problems from the real-world databases which tend to be very large, redundant, noisy, and dynamic. Each of these problems has been addressed to some extent within the machine learning. Handling some of these problems while producing useful knowledge efficiently and effectively is the main focus of this paper. In this paper, for handling some of these problems, the main objective in our study is to exploit a theory of rough sets in discovery of deterministic and nondeterministic decision rules from databases.

Key Words: Knowledge Discovery, Rough Set Theory, Deterministic Decision Rules, Nondeterministic Decision Rules, Database.

* Teacher- Basic Science Department – Faculty IT- Damascus University

1. مقدمة:

في السنوات الأخيرة، أصبحت هناك حاجة ملحة لاستخراج المعلومات من القيم غير المؤكدة وغير الدقيقة والغامضة بالتوازي مع القيم الدقيقة الموجودة في قواعد المعطيات. تقدم نظرية المجموعات التقريبية طريقة جديدة لتعلم الآلة، واكتشاف المعرفة من قواعد المعطيات، والاستنتاج المنطقي من المعطيات. وتكون منهجيتها مكمِّلة للطرق الإحصائية للاستدلال. وأهمية البحث في نظرية المجموعات التقريبية تعود إلى حقيقة أنها تستخدم فقط المعلومات الداخلية، ولا تعتمد على افتراض نموذج إضافي مثلما تفعل الطرق الأخرى (نظرية الدليل، نظرية المجموعات الضبابية،...). وهذا يعني أنه بدلاً من استخدام وسطاء إضافية أخرى، يستفيد تحليل المجموعة التقريبية فقط من بنية المعطيات المعطاة. وسمة هامة أخرى لهذه النظرية، هي قدرتها على التعامل مع الضجيج، والغموض، وعدم التأكد، والمعطيات غير المتسقة، حيث أنها تعالج بنفس طريقة معالجة المعطيات المتسقة، ولا تتطلب أي طرق إضافية.

2. هدف البحث:

في الدراسات السابقة، طورت عدة طرق لاكتشاف المعرفة من قواعد المعطيات مثل طريقة ID_3 وطريقة ID_4 وطريقة ID_5 وطريقة $C_{4.5}$ وطريقة $C_{5.0}$[15] وطرق أخرى مثل CN_2 و SLIQ وخوارزمية إزالة المرشح و AQ11 و AQ15. بينما تركز هذه المقالة على اكتشاف المعرفة من قواعد المعطيات باستخدام نظرية المجموعات التقريبية حيث كان الهدف الرئيسي من دراستنا هو كيفية استغلال هذه النظرية في اكتشاف قواعد القرار المحددة وغير المحددة من قواعد المعطيات، مع مقارنة ذلك مع طريقة $C_{5.0}$ كونها الطريقة الأكثر أهميةً وشيوعاً وتطوراً وحداثةً في مجال اكتشاف المعرفة من قواعد المعطيات. حيث طورنا في هذه المقالة عدة خوارزميات جديدة لتحقيق هذا الهدف.

3. مراحل العمل:

في إنجاز هدفنا، تتألف منهجية البحث المتبعة في هذه المقالة من عدة خطوات. في الخطوة الأولى، عرضنا نمذجة قاعدة المعطيات في المجموعة التقريبية. وفي الخطوة الثانية، عرضنا أدوات المجموعة التقريبية لتحليل المعطيات. وفي الخطوة الثالثة، عرضنا توليد قواعد القرار وبشكل خاص قواعد القرار المحددة وغير المحددة، وأخيراً، في الخطوة الربعة، عرضنا مقارنة لطريقتنا مع طريقة $C_{5.0}$.

4. نمذجة قاعدة المعطيات في المجموعة التقريبية:

تُمثل المعطيات في نموذج قاعدة المعطيات العلائقي المطور من قبل Codd (1970) بجداول تدعى علاقات(Relations). ويحتوي كل عمود في الجدول قيماً لصفة محددة لكل حدوديات المعطيات، أو السطور في الجدول. وتكون كل القيم ضمن عمود معطى عناصر من مجال الصفة لذلك العمود. ومن وجهة نظر المجموعة التقريبية، تعتبر قاعدة المعطيات العلائقية نظام معلومات فيه الحقول أوالأعمدة تعنون بواسطة الصفات، والسجلات أو السطور تعنون بواسطة الأغراض، والمدخل عند العمود a والسطر x يملك القيمة a(x). وبناءً على ذلك، يمثل كل سطر في الجدول العلائقي معلومات حول غرض معين في المجموعة الشاملة U. والاختلاف هو أن الكينونات لنظام المعلومات لا تحتاج لأن تميز بواسطة صفاتها أو بواسطة ارتباطاتها بكينونات نوع آخر [2, 14].

تعريف 1: رسمياً، نظام المعلومات يمكن أن يمثل بالرباعية S={U,A,V, f}حيثUهي مجموعة منتهية غير خالية من الأغراض: U={$x_1,x_2,..,x_n$}، وAهي مجموعة منتهية غير خالية من الصفات والتي أيضاً تصنف إلى مجموعتين منفصلتين تمثلان الصفات الشرطية C وصفات القرار D حيث: $A = C \cup D$ ، و $V = \bigcup_{a \in A} V_a$ هي مجموعة غير خالية من قيم الصفات و V_a هي مجال الصفة a، و $f : U \times A \to V$ هي دالة المعلومات التي تخصص قيمة للصفة $a \in A$ ، من أجل $x_i \in U$ [7]. يوضح الجدول(1) مثالاً عن نظام المعلومات.

الجدول 1: نظام معلومات

	Condition attributes			Decision
	A	B	C	D
x_1	a1	b1	c2	d0
x_2	a1	b1	c0	d1
x_3	a1	b1	c1	d1
x_4	a0	b1	c2	d0
x_5	a0	b0	c0	d0
x_6	a0	b1	c1	d1

5. أدوات المجموعة التقريبية لتحليل المعطيات:

تنمو المعطيات المخزنة في الحاسوب بالحجم بشكل سريع جداً لدرجة أننا قد نغرق بفيضان من المعلومات الحاسوبية. سواءً في التطبيقات العلمية أو تطبيقات الأعمال، آلاف من قواعد المعطيات طورت لتخزين المعطيات الملائمة لصف نشاطات محدد. وهذه المعطيات المجمعة في قواعد المعطيات، غالباً ما تعاني من بعض المشكلات الخاصة بطبيعة المعطيات المجمعة التي تتجه لأن تكون غير دقيقة وغامضة وأحيانا غير متسقة. تقدم نظرية المجموعات التقريبية الصياغة الضرورية لإجراء تحليل المعطيات من المعطيات المبهمة وغير الدقيقة. كما إنها تقدم تقنيات للتعامل مع المفاهيم الغامضة، وتحليل تبعية المعطيات، وقياس أهمية المعطيات، وتقنيات لتحديد الصفات الملائمة من مجموعة السمات المعطاة.

5-1 الغموض وتقريب المجموعة:

تدرك الأغراض في المجموعات التقريبية فقط من خلال المعلومات المتاحة حولها. ومن جهة أخرى، في التطبيقات الحياتية الحقيقية، ليس من الممكن دائماً الحصول على معلومات كاملة تعرِّف الأغراض في عالمنا.

يقدم عدم الكمال للمعلومات في مجموعة معطيات معطاة مصدراً ما للغموض وعدم الدقة في مجموعة المعطيات. وهذا يمثل مصدراً لعدم التأكد في المعرفة المكتسبة من هذه المعطيات. ويفترض أن تتمذج المعرفة المكتسبة من مجموعات المعطيات بواسطة مفاهيم، كل واحد منها يمثل بواسطة مجموعة من الأغراض. ومن جهة أخرى، إذا كانت المعرفة غير مؤكدة، عندئذ المفاهيم المناظرة يمكن أن تكون غامضة وعندها يمكن أن تمثل بمجموعات تقريبية.

في منهج المجموعات التقريبية، يكون الغموض بسبب نقص المعلومات التي تصف الأغراض في المجموعة الشاملة. وتحل مشكلة الغموض باستخدام مفاهيم تقريب المجموعة. فكرة الغموض، كما وصفت سابقاً، تكون مرتبطة بما يدعى الرؤية الحدية التي أول من قدمها [1] Frege. والمفهوم يكون غامضاً إذا كانت هناك بعض الأغراض التي لا نستطيع تصنيفها فيما إذا كانت من المفهوم أو من متممته. ومثل هذه الأغراض ستقع في المنطقة الحدية لذلك المفهوم.

يستبدل المفهوم الغامض في المجموعة التقريبية بزوج من المفاهيم الدقيقة تدعى تقريباته العلوية والسفلية. ويتألف التقريب السفلي لمفهوم من كل الأغراض التي تنتمي

بالتأكيد إلى المفهوم. ويتألف التقريب العلوي لمفهوم من كل الأغراض التي من الممكن أن تنتمي إلى المفهوم. ويشكل الفرق بين التقريب السفلي والتقريب العلوي ما يدعى المنطقة الحدية للمفهوم الغامض.

تعريف 2: ليكن S={U,A,V,f} نظام معلومات. إذا أخذنا $P \subset A$ ، وليكن x_i،x_j غرضان من U، نعرف العلاقة الثنائية IND، التي تدعى علاقة عدم إمكانية التمييز، على مجموعة الصفات P− حيث P هي مجموعة الصفات المتساوية لكلا الغرضين− كما يلي[13,7]:

$$IND(P) = \{(\chi_i, \chi_j) \in U \times U \mid \forall a \in P; f(\chi_i, a) = f(\chi_j, a)\} \qquad (1)$$

إن IND(P) هي علاقة تكافؤ تقسم المجموعة الشاملةU إلى مجموعة من صفوف التكافؤ. هذه المجموعة سيشار لها بالتصنيف ويرمز لها بـ U/IND(P). العناصر التي تنتمي إلى نفس صف التكافؤ تكون غير قابلة للتمييز، وخلاف ذلك تكون العناصر قابلة للتمييز بالنسبة لـ P. وبناءً على ذلك من أجل كل $a \in P$ إذا كان $f(x_i,a)=f(x_j,a)$، عندها بإمكاننا القول بأن x_iوx_j غير قابلين للتمييز بواسطة مجموعة الصفات P في نظام المعلومات S.

تعريف 3:التقريب السفلي أو الحد السفلي لـ X، يرمز له بـ $\underline{Apr}(X)$ ، هو اجتماع كل المجموعات الأولية التي كل واحدة منها يمكن أن تصنف بأنها تنتمي بالتأكيد إلى المجموعة X اعتماداً على معلومات التصنيف الممثلة بواسطة علاقة عدم إمكانية التمييز IND[7]:

$$\underline{Apr}(X) = \{x_i \in U \mid [x_i]_{IND} \subset X\} \qquad (2)$$

تعريف 4: التقريب العلوي أو الحد العلوي لـ X، يرمز له بـ $\overline{Apr}(X)$ ، هو اجتماع كل المجموعات الأولية التي كل واحده منها يمكن أن تصنف بأنها من الممكن أن تنتمي إلى المجموعة X اعتماداً على معلومات التصنيف الممثلة بواسطة علاقة عدم إمكانية التمييز IND[7]:

$$\overline{Apr}(X) = \{x_i \in U \mid [x_i]_{IND} \cap X \neq \phi\} \qquad (3)$$

تعريف 5:المنطقة الحدية للمجموعة X، يرمز لها بـ BND(X)، هي اجتماع كل المجموعات الأولية التي من أجل كل منها لا يمكن أن نحدد بالتأكيد فيما إذا كانت الأغراض في هذه المنطقة تنتمي إلى X أو إلى متممتها X–[7]:

$$BND(X) = \overline{Apr}(X) - \underline{Apr}(X) \qquad (4)$$

تظهر الأشكال من الشكل (1) حتى الشكل (3) الخوارزميات البسيطة لحساب التقريبات العلوية والتقريبات السفلية لـ P والمنطقة الحدية لمفهوم محدد C بالنظر إلى المعلومات المتاحة المعرفة بواسطة مجموعة الصفات الشرطية P.

يمكن أيضاً أن يوصف الغموض عددياً بتعريف المعامل التالي (معامل الدقة)[7]:

$$\alpha_p(X) = \frac{|P - Lower - Approx(X)|}{|P - Upper - Approx(X)|} \qquad (5)$$

حيث X هي مجموعة الأغراض التي تعرف مفهومنا، و|X| يرمز إلى رئيسي المجموعة X. ومن الواضح أن $0 \le \alpha_p(X) \le 1$. فإذا كان $\alpha_p(X) = 1$ عندئذ المجموعة X تكون مجموعة واضحة بالنسبة للمعلومات المتاحة P، وعندها يكون المفهوم C دقيقاً بالنسبة لـP. وفي الحالة الأخرى، إذا كان $\alpha_p(X) < 1$ عندئذ المجموعة X تكون تقريبية والمفهوم C يكون غامضاً بالنسبة لـ P. لاحظ أن درجة الغموض للمفهوم تتفاوت اعتماداً على قيمة $\alpha_p(X)$ من المفهوم الغامض كلياً عندما $\alpha_p(X) = 0$ إلى المفهوم الغامض جزئياً عندما $0 < \alpha_p(X) < 1$.

```
Algorithm (P Lower Approximation of Concept C)

Input:
(a) The task relevant data relation
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
(d) The Concept C
Output:
The set of objects P-LowerApprox(X) defining the P Lower Approximation of C
Method
        {Collect the set of objects defining concept C in set X
        Compute complement of set X with respect to R in set NotX
        Initially P-LowerApprox(X) = X
        For each object I in X
           For each object J in NotX
           If (object I and object J are indiscernible with respect to P)
           Then remove I from P-LowerApprox(X)
        return(P-LowerApprox(X))}
```

الشكل 1: خوارزمية لحساب التقريب السفلي للمفهوم C

Algorithm (P Upper Approximation of Concept C)

Input:
(i) The task relevant data relation R
(ii) The set P defining condition attributes
(iii) The set Q defining decision attributes, usually containing one attribute.
(iv) The Concept C
Output:
The set of objects P-UpperApprox(X) defining the P Upper Approximation of C
Method

```
Collect the set of objects defining concept C in set X
Compute complement of set X with respect to R in set NotX
Initially P-UpperApprox(X) = X
For each object I in X
        For each object J in NotX
        If (object I and object J are indiscernible with respect to P)
        Then add J to set P-UpperApprox(X)
return(P-UpperApprox(X))
```

الشكل 2: خوارزمية لحساب التقريب العلوي للمفهوم C

Algorithm (P Boundary Region of Concept C)

Input:
(i) The task relevant data relation R
(ii) The set P defining condition attributes
(iii) The set Q defining decision attributes, usually containing one attribute.
(iv) The Concept C
Output:
The set of objects Boundary defining the P Boundary Region of Concept C
Method

```
Collect the set of objects defining concept C in set X
Compute complement of set X with respect to R in set NotX
Initially P-Bound(X) is an empty set
For each object I in X
        For each object J in NotX
        If (object I and object J are indiscernible with respect to P)
        Then add objects I and J to set P-Bound(X)
return(P-Bound(X))
```

الشكل 3: خوارزمية لحساب المنطقة الحدية للمفهوم C

وبناءً على ذلك، استخدام معامل الدقة في سياق المجموعات التقريبية، يمكننا من تحديد طبيعة أي مفهوم بالنظر إلى المعلومات المتاحة. والمفهوم يكون إما دقيقاً أو

غامضاً كلياً أو جزئياً اعتماداً على المعلومات المتاحة سواءً كانت كافية لتعريف المفهوم أو غير كافية وإلى أي مدى نستطيع أن نعرفه بدقة. يظهر الشكل (4) خوارزمية بسيطة لتحديد طبيعة مفهوم محدد C بمعرفة الصفات الشرطية P.

2-5 المنطقة الموجبة وتبعية المعطيات:

توجد بعض الارتباطات المهمة التي تظهر في المعطيات المخزنة في قواعد المعطيات. وهذا الارتباط يمكن أن يأخذ شكل التبعية الدالية أو التبعية الدالية الجزئية بين المعطيات. ولتوصيف أو اكتشاف هكذا تبعية بين المعطيات يمكن أن نستخدم طرق نظرية المجموعات التقريبية.

في منهج المجموعات التقريبية، كشف التبعية بين الصفات هو المرحلة الأولية والابتدائية في تحليل المعطيات. في هذا المقطع سنناقش الأفكار الهامة للمنطقة الموجبة وتبعية الصفات في نظرية المجموعات التقريبية.

```
Algorithm (Identifying Nature of Concept C)

Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
(d) The Concept C
Output:
NofC defining Nature of Concept C viewed through information P
Method
        Compute P-LowerApprox(X)
        CardLower = Cardinality of P-LowerApprox(X)
        Compute P-UpperApprox(X)
        CardUpper = Cardinality of P-UpperApprox(X)
        AccurCoef(X) = CardLower / CardUpper
        If (AccurCoef(X) == 1)
        Then NofC = Precise
        Else If (AccurCoef(X)> 0)
        Then NofC = Par-Vague        // partially vague concept
        Else NofC = T-Vague          // totally vague concept
        return(NofC)
```

الشكل 4: خوارزمية لتحديد طبيعةالمفهوم C

بمعرفة علاقة قاعدة المعطيات R الملائمة لمهمة محددة، لتكن R تمثل المجموعة الشاملة U، و A مجموعة الصفات في R، و $P,Q \subseteq A$. المنطقة الموجبة للتصنيف U/IND(Q)بالنسبة لمجموعة الصفات P، هذا يعني المنطقة الموجبة-P لـ Q، تعرف كما يلي [7]:

$$POS(Q) = \bigcup_{X \in U/IND(Q)} \underline{P}X \quad (6)$$

تحتوي المنطقة الموجبة–P لـ Q كل الأغراض التي بمعرفة المعلومات المتاحة الممثلة في مجموعة الصفات P، يمكن أن تصنف بالتأكيد إلى صف واحد من صفوف التصنيف IND(Q). يظهر الشكل (5) خوارزمية بسيطة لحساب المنطقة الموجبة لصفة قرار محددة Q بمعرفة المعلومات المتاحة والمعرَّفة بواسطة مجموعة الصفات الشرطية P.

```
Algorithm  (P Positive Region of Q)
Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
Output: P-Pos(Q)  defining  the P Positive Region of Q
Method
        Initially P-Pos(Q) = all objects in R
        For each object I in R
          For each object J  in R
          If objects I and J are discernible with respect to Q
             If objects I and J are indiscernible with respect to P
              Remove objects I and J  from P-Pos(Q)
        return(P-Pos(Q))
```

الشكل 5: خوارزمية لحساب المنطقة الموجبة P لـ Q

تستخدم فكرة المنطقة الموجبة في المجموعات التقريبية لقياس التبعية بين الصفات الشرطية وصفة القرار في علاقة قاعدة المعطيات. وبناءً على ذلك، لقياس تبعية صفة القرار Q على مجموعة الصفات الشرطية P، سنستخدم حجم المنطقة الموجبة–Pلـ Q كما يلي[7]:

$$\gamma_P(Q) = \frac{Card(POS_P(Q))}{Card(U)} \qquad (7)$$

الصيغة $\gamma_P(Q)$ تعرف بعامل التبعية لـ Q على المجموعة P. هذا يعني، إننا نقول أن Q تابعة لـ P بدرجة $\gamma_P(Q)$. ويتراوح معامل التبعية من القيمة صفر التي تشير لعدم وجود تبعية لـ Q على P إلى القيمة واحد التي تشير إلى التبعية الكلية. ويشير معامل التبعية بين الصفر والواحد إلى التبعية الجزئية التي ستكون أكبر من صفر

وأصغر من واحد. يظهر الشكل (6) خوارزمية لقياس التبعية لصفة قرار محددة Q على مجموعة الصفات الشرطية المتاحة P في علاقة قاعدة معطيات معطاة.

```
Algorithm (P Positive Region of Q)
Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
Output: P-Pos(Q) defining the P Positive Region of Q
Method
        Initially P-Pos(Q) = all objects in R
        For each object I in R
          For each object J in R
          If objects I and J are discernible with respect to Q
             If objects I and J are indiscernible with respect to P
              Remove objects I and J from P-Pos(Q)
        return(P-Pos(Q))
```

الشكل 6: خوارزمية لقياس تبعية المعطيات

3-5 اختزال مجموعة المعطيات العلائقية:

إن إحدى المشكلات الرئيسية التي يمكن حلها باستخدام منهج المجموعة التقريبية، هي مختزل الصفات في علاقة قاعدة المعطيات، أو كما تسمى في سياق المجموعة التقريبية، نظام المعلومات. وطرق المجموعة التقريبية يمكن استخدامها في تحديد عناصر أو سمات المعطيات التي لا تقدم مساهمة معنوية في تعريف مفهوم محدد، ولذلك بإمكانها أن تساعدنا في التعامل مع المشكلات المتعلقة بفائض المعلومات بإزالة المعلومات الزائدة.

في اختزال علاقة المعطيات، يلعب الدور الرئيسي مفهومان أساسيان للمجموعات التقريبية واللذان يمكناننا من الحصول على مختزل الصفات بدون خسارة في قوة التصنيف لنظام المعلومات الأصلي وهذان المفهومان هما المختزل واللب[13].المختزل لعلاقة قاعدة المعطيات، أو لنظام المعلومات، كما وصف من قبل، هو جزءها الأساسي الذي يكون كافياً لتمييز كل المفاهيم الأساسية الظاهرة في علاقة المعطيات المأخوذة بعين الاعتبار. اللب هو الجزء الأعم والأكثر أهمية لكل المختزلات في نظام المعلومات المتناول. ويتطلب اختزال علاقة المعطيات تحديد الصفات الزائدة والصفات الأساسية في نظام المعلومات النشط. وإذا كانت الصفة زائدة في نظام معلومات عندئذ يمكن إزالتها

من نظام المعلومات بدون تغيير في علاقة التبعية للنظام الأصلي. ومن جهة أخرى، تحمل الصفة غير الزائدة معلومات أساسية حول أغراض نظام المعلومات، ولذلك يجب أن نحافظ عليها إذا أردنا الحفاظ على التبعية في نظام المعلومات الأصلي.

5-3-1 اللب ومصفوفة إمكانية التمييز:

اللب هو أحد المفاهيم الأكثر أهمية في نظرية المجموعات التقريبية. واللب يشمل الصفات غير الزائدة لتمييز الأغراض في نظام المعلومات. ولذلك تكون الصفات في اللب محتواه في كل المختزلات. وبناءً على ذلك، غالباً ما يستخدم اللب كقاعدة لحساب المختزل.

إمكانية التمييز هي فكرة هامة أخرى في منهج المجموعات التقريبية لتحليل المعطيات. وإمكانية تمييز الأغراض في نظام المعلومات غالباً ما تكون أكثر أهمية من القيم المحددة للسمات. وهكذا في الغالب يمثل نظام المعلومات بمصفوفة إمكانية التمييز. ومصفوفة إمكانية التمييز هي مصفوفة مربعة لها أسطر وأعمدة تعنون بالأغراض. وكل مدخل للمصفوفة هو مجموعة جزئية من الصفات التي تميز الأغراض.

تعريف 6:مصفوفة إمكانية التمييز لمجموعة الصفات $P \subseteq A$ في نظام معلومات S(U, A) يمكن أن توصف بالصيغة التالية[13,7]:

$$D_{\chi_i,\chi_j} = \{a \in A: f(\chi_i, a) \neq f(\chi_j, a) \; where \; \chi_i \; and \; \chi_j \in U\} \qquad (8)$$

مصفوفة إمكانية التمييز المناظرة لنظام المعلومات الموضح بالجدول (1) تعرض أدناه:

	x1	x2	x3	x4	x5	X6
x1						
x2	{C}					
x3	{C}	{C}				
x4	{C}	{A,C}	{A,C}			
x5	{A,B,C}	{A,B}	{A,B,C}	{B,C}		
x6	{A,C}	{A,C}	{A}	{C}	{B,C}	

من المصفوفة السابقة نلاحظ أن الصفتان (A) و(C) معاً تكونان كافيتان لتمييز كل الأغراض في نظام المعلومات. وبناءً على ذلك، تكون المجموعات الأولية لعلاقة عدم إمكانية التمييز المعرفة بواسطة المجموعة {A, C} وحيدة العنصر، بمعنى المجموعات: {x1}, {x2}, {x3}, {x4}, {x5}, {x6} وكذلك تكون المجموعات الأولية لعلاقة عدم إمكانية التمييز المعرفة بواسطة المجموعة المؤلفة من كل الصفات الثلاث وحيدة العنصر، وهكذا الصفة (B) تعتبر صفة زائدة في نظام المعلومات.

يرتبط اللب بشكل قوي مع مصفوفة إمكانية التمييز. وباستخدام مصفوفة إمكانية التمييز، نستطيع بسهولة حساب اللب لنظام المعلومات. يظهر الشكل (7) خوارزمية بسيطة لحساب اللب النسبي Q لعلاقة مجموعة المعطيات الملائمة للمهمة.

```
Algorithm (Q-Relative Core of an I.S)

Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
Output:
The set of attributes Core_Q(P) Q-relative Core of P
Method
        Initially Core_Q(P) = empty set
        While I,J go through all pairs of objects in R
           if (object I and object J are discernible with respect to Q)
           if (object I and object J are discernible on a single attribute p in P)
           Core_Q(P) = Core_Q(P) + p
        return (Core_Q(P))
```

الشكل 7: خوارزمية لحساب اللب النسبي Q للمجموعة P

5-3-2 أهمية الصفات في المعطيات:

في نموذج قاعدة المعطيات، عادة كل الصفات نتعامل معها بأهمية متساوية وهذا ليس دائماً صحيح في العديد من تطبيقات الحياة الحقيقة. والصفات المختلفة غالباً ما تلعب أدوار مختلفة في تحديد علاقة التبعية بين الشروط والقرارات في علاقة مجموعة المعطيات.

تمثل أهمية الصفة المفردة (a) في مجموعة الصفات الشرطية P بالنسبة للتبعية بين pوQ، بمعامل الأهمية $\mu_{P,Q}(a)$ والذي يمكن تعريفه بالصيغة التالية [7]:

$$\mu_{P,Q}(a) = \frac{Card(POS_P(Q)) - Card(POS_{P-\{a\}}(Q))}{Card(U)} \qquad (9)$$

$\mu_{P,Q}(a)$ يعرِّف أهمية $a \in P$ بالنسبة لصفة القرار Q. أحياناً يتطلب قياس أهمية الصفة a في مجموعة الصفات الشرطية P أن يحسب معامل الأهمية

$\mu_{P,Q}(a)$ بالنسبة للتصنيف الأصلي IND(A). وفي هذه الحالة أهمية $a \in P$ يرمز لها بـ $\mu_P(a) = \mu_{P,A}(a)$.يظهر الشكل (8) خوارزمية لقياس الأهمية لصفة محددة $a \in P$ بالنسبة لصفة قرار محددة Q.

3-3-5 اختزال الصفات في نظام المعلومات:

تصف مجموعة الصفات في نظام المعلومات الأغراض المهتم بها. ما يحدث أن مجموعة الصفات تقدم معلومات حول الأغراض أكثر مما نحتاجه لتمييزها. وفي هكذا حالة، بعض الصفات من الممكن إزالتها بدون خسارة في المقدرة على تصنيف الأغراض. تمكننا فكرة المختزلات في سياق المجموعات التقريبية من الاختزال المثالي للصفات بدون خسارة في قوة التصنيف في نظام المعلومات[11]. ومختزل الصفات هو المجموعة الجزئية التي بإمكانها تمييز كل الأغراض في نظام المعلومات مثلما تفعل مجموعة الصفات بأكملها. يظهر الشكل (9) خوارزمية لإيجاد المختزل لمجموعة الصفات في علاقة مجموعة المعطيات بالنسبة لصفة قرار محددة Q.

Algorithm(Significance of attribute a∈P relative to Q)

Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
(d) Attribute a

Output:
$SigRel_{P,Q}(a)$ defining the significance of attribute a∈P relative to Q

Method

```
Compute P-Pos(Q)
P* = P
Remove attribute a from set P*
Compute P*-Pos(Q)
TempPos = P-Pos(Q) - P*-Pos(Q)
CardTempPos = cardinality of TempPos
CardR = cardinality of the data set relation R
SigRel_P,Q(a) = CardTempPos / CardR
return(SigRel_P,Q(a))
```

الشكل 8: خوارزمية لقياس أهمية المعطيات

Algorithm (Q Relative Reduct of P)

Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.

Output:
The set of attributes $Red_Q(P)$ representing Q-relative Reduct of P

Method
Compute $Core_Q(P)$
Initially $Red_Q(P) = Core_Q(P)$
$P' = P - Red_Q(P)$
For each $a \in P$
Compute significant value of attribute (a) $SigRel_{P,Q}(a)$
Sort the set of attributes in P' based on $SigRel_{P,Q}(a)$ for each $a \in P'$
While $Depend_P(Q) \neq Depend_{Red(P,Q)}(Q)$
{Select an attribute a with the highest $SigRel_{P,Q}(a)$ in P'
If $a \notin Red_Q(P)$ Then $Red_Q(P) \cup \{a\}$
Compute $Depend_{Red(P,Q)}(Q)$ }
return($Red_Q(P)$)

الشكل 9: خوارزمية لحساب المختزل النسبي Q للمجموعة P

6 توليد قواعد القرار:

يعتبر توليد قواعد القرار أحد أهم المشكلات التي يمكن أن تقرب باستخدام نظرية المجموعات التقريبية[10]. وفي سياق اكتشاف المعرفة، عادة مجموعة معطياتنا تكون قاعدة معطيات علاقية.

تعتبر قاعدة المعطيات العلائقية، كما ذكرنا سابقاً، كنظام معلومات. كل سطر في نظام المعلومات هذا(الجدول العلاقي) يمثل معلومات حول غرض أو حالة محددة من المجموعة الشاملة. ومجموع كل الحدوديات في مجموعة المعطيات المأخوذة بعين الاعتبار تمثل مجموعة أمثلة التدريب. وفي عملية توليد القاعدة، الصفات المأخوذة بعين الاعتبار تصنف في مجموعتين منفصلتين: الصفات الشرطية وصفات القرار، وعادة صفة قرار فردية تكون كل ما هو مطلوب. وكنتيجة لهذا تصنيف للصفات، نظام المعلومات يمكن أن يفهم كجدول قرار.

قاعدة القرار في نظام المعلومات أو في قاعدة المعطيات هي تركيب من القيم لبعض الصفات الشرطية مطبق إلى قيمة محددة لصفة القرار:

$$r:Condition(r) \rightarrow Decision(r)$$

وعادة، تكون مدخلات عملية توليد القاعدة مجموعة الصفات الشرطية P وصفة القرار Q. وفي هكذا حالة، الهدف هو توليد القواعد التي تعرِّف المفاهيم المختلفة تحت صفة القرار تلك. ومن جهة أخرى، أحياناً ما يطلب توليد القواعد التي تعرف مفهوم محدد يظهر في Q. وفي هكذا حالة، المفهوم المطلوب يجب أن يحدد كمدخل في عملية التعلم.

المفهوم في جدول علاقي هو ببساطة قيمة محددة V_Q لصفة القرار Q. ومن ثم الهدف لمهمة التعلم سيكون توليد القواعد التي تعرِّف مفهوم محدد يظهر في Q. وهذا يعني، مجموعة الأغراض التي لها القيمة V_Q لصفة القرار Q. وفي هكذا حالة، مجموعة الأغراض التي تملك القيمة V_Q لصفة قرار مؤكدة Q ستشير إلى المفهوم الهدف أو إلى مجموعة الحالات المتاحة التي تعرف المفهوم V_Q. والقواعد المولدة التي تعرِّف هذا المفهوم ستكون من الشكل[7]:

$$r: (a_{i1}=v_{i1}) \wedge (a_{i2}=v_{i2}) \wedge .. \wedge (a_{in}=v_{in}) \rightarrow Q=v_Q \qquad (10)$$

حيث أن $a_{i1},a_{i2},..,a_{in}$ هي مجموعة جزئية من مجموعة الصفات الشرطية P. وv_Q هو المفهوم الهدف الذي يظهر في صفة القرار Q.

قياس قوة القاعدة يقدم لكل قاعدة من قواعد القرار المولدة. وباستخدام هذا القياس، نكون قادرين على التفريق بين القواعد القوية والقواعد الضعيفة. ونقصد بالقواعد القوية تلك القواعد التي تعكس الأنماط المتكررة جداً التي تظهر في المعطيات. والقواعد القوية لا يفترض بالضرورة أن تكون قواعد دقيقة أو مؤكدة، ومن جهة أخرى، يجب أن تكون صحيحة أو على الأغلب صحيحة وتعكس درجة ما من التبعية في مجموعة المعطيات. S(r) تعرِّف قوة القاعدة (r)، والتي تعرَّف بعدد الأغراض في علاقة مجموعة المعطيات الأصلية التي قيم صفاتها تطابق القاعدة (r). وقوة القاعدة تقدم لكل قاعدة من القواعد المولدة كقياس لدرجة الاعتماد على هذه القاعدة من بين القواعد الأخرى المولدة.

6-1 قواعد القرار في نظام المعلومات:

تمثل القواعد التبعية في مجموعة المعطيات التي منها تولد. وهي أيضاً تمثل المعرفة المستخرجة التي يمكن استخدامها إما لتوصيف مفهوم محدد أو لتصنيف أغراض جديدة ليست موجودة في مجموعة المعطيات الأصلية لأحد صفوف القرار.

مختزل نظام المعلومات، كما وصف سابقاً، هو المجموعة الجزئية الصغرى للصفات المهتم بها التي تملك نفس القوة للتمييز بين الصفوف المختلفة في صفة القرار مثل ما تفعل كل الصفات في نظام المعلومات أو مجموعة معطيات الهدف.

يعتمد توليد القواعد في سياق المجموعات التقريبية بشكل أساسي على اختزال الصفات الزائدة في مجموعة معطيات الهدف[4]. وهكذا عندما تحسب المختزلات، فإن مهمة إنتاج قواعد القرار لصفة قرار محددة، في نظام المعلومات، تكون قد تمت تقريباً. لذلك لتحويل مختزل نسبي إلى قاعدة، ببساطة نربط قيم الصفات الشرطية لصف الغرض الذي منه حُسب المختزل إلى صفات المختزل المناظرة. ومن ثم صفة القرار التي تشكل الجزء الناتج للقاعدة تضاف. يظهر الشكل (10) خوارزمية بسيطة لتوليد قاعدة القرار. في هذه الخوارزمية، اهتممنا فقط بتوليد قواعد القرار التي تغطي علاقة مجموعة المعطيات المعطاة. لذلك، لا يجب أن نميز بين الصفوف التي تطبق إلى نفس القرار.

2-6 القرارات المحددة مقابل القرارات غير المحددة:

القواعد المولدة من أجل صفة قرار محددة، في علاقة مجموعة المعطيات، تكون إما متسقة، أي، "قواعد محددة " أو تكون غير متسقة " قواعد غير محددة ". والقواعد المحددة تحدد قرارات فريدة بينما القواعد غير المحددة تحدد مجموعة من القرارات الممكنة التي يمكن أن تُصنع على أساس الشروط المعطاة.

اعتماداً على مفاهيم تقريب المجموعة والمنطقة الموجبة وقياس التبعية $\gamma_C(D)$ في نظرية المجموعة التقريبية، علاقة مجموعة المعطيات أو جدول القرار يمكن أن يجزأ أفقيا إلى جدولين جزئيين أحدهما محدد أي $\gamma_C(D)=1$، والآخر غير محدد كلياً أي $\gamma_C(D)=0$. في هذه المقالة سنهتم فقط بتوليد القواعد المحددة والقواعد غير المحددة.

Algorithm (Decision Rule Generation)
Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
Output:
The set of generated decision rules for R
Method
{ T = all objects in R // T is our target data set
Compute $Red_Q(P)$
Reduce T by removing attributes not in $Red_Q(P)$
Merge similar objects
Transform the reduced T to logical rules
Calculate strength for each rule
return(N)} // N defines the number of generated rules

الشكل 10: خوارزمية لتوليد قاعدة قرار

6-2-1 توليد القاعدة المحددة:

القواعد المحددة لصفة قرار محددة Q هي القواعد التي بالنظر إلى مجموعة الصفات الشرطية المتاحة تحدد بشكل فريد قرارات الصفة Q. وهكذا باستخدام القواعد المحددة نستطيع أن نميز بين المفاهيم المختلفة التي تظهر في صفة القرار Q.

علاقة مجموعة المعطيات، كما ذكرنا سابقاً، يمكن أن تقسم إلى جزأين مختلفين، الأول يكون محدد كلياً والآخر غير محدد. الجزء الأول من العلاقة سيكون مجموعة معطيات الهدف الخاصة بنا التي منها ستولد قواعد القرار المحددة.

Algorithm (Deterministic Rule Generation)
Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
Output:
The set of deterministic decision rules for R
Method
{Compute $POS_P(Q)$
T = objects in $POS_P(Q)$ // T represent target data set
Compute $Red_Q(P)$
Reduce T by removing attributes not in $Red_Q(P)$
Merge similar objects
Transform the reduced T to logical rules
Calculate strength for each rule
return(N)} // N defines the number of generated rules

الشكل 11: خوارزمية لتوليد القاعدة المحددة

تقدم نظرية المجموعات التقريبية أداة هامة التي بواسطتها نتمكن من تعريف الجزء المحدد لعلاقة مجموعة المعطيات، أي المنطقة الموجبة. والمنطقة الموجبة، بالتعريف، تحتوي كل الأغراض في علاقة مجموعة المعطيات التي يمكن أن تصنف في صفوف مختلفة لصفة القرار المختارة بدون خطأ اعتماداً على معلومات التصنيف المتاحة، أي، مجموعة الصفات الشرطية. وبناءً على ذلك، الاستراتيجية الأساسية في خوارزميتنا لتوليد القاعدة المحددة تعتمد على تطبيق مفهوم المنطقة الموجبة على علاقة مجموعة المعطيات الأصلية وتصفية تلك الحدوديات المعرِّفة للأغراض التي تصنَّف بشكل وحيد إلى قراراتها عن الأغراض الأخرى التي تصنف إلى أكثر من قرار واحد. والخوارزمية لتوليد القاعدة المحددة تم تقديمها في الشكل (11).

سنوضح عمل خوارزميتنا السابقة كما يلي: ليكن لدينا جدول المعطيات (2).

الجدول 2: جدول معطيات

U	A	B	C	D
x1	a0	b1	c0	d1
x2	a1	b0	c0	d1
x3	a1	b1	c1	d1
x4	a0	b1	c2	Do
x5	a1	b0	c0	Do
x6	a0	b1	c1	d1

نلاحظ أن المجموعات الابتدائية المتولدة عن علاقة عدم إمكانية التمييز (U/C) هي:

{x1} , {x2 , x5} , {x3} , {x4} , {x6}

إن التقريب السفلي للمفهوم D هو: $\underline{Apr}(D)$={x1, x3, x6} والتقريب العلوي للمفهوم D هو: $\overline{Apr}(D)$ ={x1, x2, x3, x5, x6}

إن التقريب السفلي للمفهوم notD هو: $\underline{Apr}(notD) = \{x_4\}$ والتقريب العلوي للمفهوم notD هو:

$\overline{Apr}(notD) = \{x_2, x_4, x_5\}$. ومنه تكون المنطقة الموجبة لصفة القرار D هي:

$$POS_C(D) = \bigcup_{X \in U/D} \underline{Apr}(X) = \{x1, x3, x4, x6\}$$

وبعد إيجاد المنطقة الموجبة يجب أن نقوم بحساب مختزل الصفات الشرطية باستخدام مصفوفة إمكانية التمييز. وبالحساب يمكن أن نلاحظ أن نظام المعلومات المعطى يملك أكثر من مختزل له نفس القوة في التصنيف، وأحد هذه المختزلات هو الموضح بالجدول(3).

الجدول 3: مختزل

U	A	C	D
x1	a0	c0	d1
x2	a1	c0	d1
x3	a1	c1	d1
x4	a0	c2	Do
x5	a1	c0	Do
x6	a0	c1	d1

ومنه القواعد المحددة التي ستنتج معنا بعد حساب المنطقة الموجبة وإيجاد المختزل للسمات الشرطية ستكون من الشكل:

$$r : Condition(r) \xrightarrow{s(r)} Decision(r)$$

حيث S(r) يمثل قوة القاعدة المولدة والتي تمثل عدد الأمثلة الموجبة في قاعدة المعطيات الأصلية، وبذلك تكون القواعد المحددة المولدة لمثالنا السابق هي:

$$r1:(A=a0)\wedge(C=c0)\xrightarrow{1}(D=d1)$$
$$r2:(A=a1)\wedge(C=c1)\xrightarrow{1}(D=d1)$$
$$r3:(A=ao)\wedge(C=c2)\xrightarrow{1}(D=d0)$$
$$r4:(A=a0)\wedge(C=c1)\xrightarrow{1}(D=d1)$$

6-2-2 توليد القاعدة غير المحددة:

تقليدياً، معظم العمل المنجز في توليد قاعدة القرار يركز على توليد القواعد المحددة لأنها تعرف مجموعة المعطيات التي تكون متسقة كلياً. إلا أن، مجموعة معطيات العالم الحقيقي غالباً ما تعاني من بعض عدم الاتساق في المعطيات المجمعة. ويعني عدم الاتساق بأن هناك على الأقل غرضان بنفس التوصيف، أي "نفس القيم الشرطية " لكن تطبق إلى قرارات مختلفة[4]. وهذه المشكلة عادة تشير أن التوصيف من المحتمل أن يكون غير كافياً لتمييز الأغراض في مجموعة المعطيات هذه. كنتيجة لمثل عدم الاتساقات هذه في المعطيات، من غير الممكن دائماً الحصول على مجموعة القواعد المحددة التي تغطى بشكل كامل مجموعة معطيات التدريب بأكملها.

القواعد غير المحددة، كما ذكر سابقا، هي القواعد التي تحدد مجموعة القرارات الممكنة اعتماداً على الشروط المتاحة. يظهر الشكل (12) خوارزمية لتوليد مجموعة القواعد غير المحددة من علاقة مجموعة المعطيات المعطاة. الهدف من الخوارزمية أن تكون قادرة على إيجاد التبعية الهامة في مجموعة المعطيات، حتى بوجود عدم الاتساق والغموض فيها. عدم الاتساق والغموض في مجموعة المعطيات يمكن أن ينتجان من المعلومات غير الكافية. هناك حالات حيث أن بعض قيم الصفات تكون مفقودة لأغراض محددة في مجموعة المعطيات. وقد يحدث أيضاً أن الصفات الملائمة للقرار لا تكون موجودة في مجموعة المعطيات المتوفرة.

الوجود لمثل عدم الاتساقات هذه والغموض في مجموعة المعطيات قد يسبب في توليد القواعد المحددة التي تكون خاصة جداً لمجموعة المعطيات. ولذلك باستخدام هكذا قواعد في عملية التصنيف، من الممكن أن يكون هناك العديد من الأغراض التي لم تغطى بأي قاعدة من القواعد المولدة. وكنتيجة لذلك، قد نجد أنفسنا مجبرين على استخدام القواعد غير المحددة في هكذا حالات.

Algorithm (Non Deterministic Rule Generation)

Input:
(a) The task relevant data relation R
(b) The set P defining condition attributes
(c) The set Q defining decision attributes, usually containing one attribute.
Output:
The set of non deterministic decision rules for R
Method
{Compute $POS_P(Q)$
T = R - $POS_P(Q)$ // T represent target data set
Compute $Red_Q(P)$
Reduce T by removing attributes not in $Red_Q(P)$
Merge similar objects
Transform the reduced T to logical rules
Calculate strength for each rule
Calculate certainty for each rule
return(N)} // N defines the number of generated rules

الشكل 12: خوارزمية لتوليد القاعدة غير المحددة

القواعد غير المحددة هي القواعد التي تحدد القرارات التي تكون ممكنة، ولذلك من الهام أن تزود كل قاعدة من هذه القواعد بمقياس للتأكد. وهكذا بالإضافة إلى المقياس المقدم في حالة القواعد المؤكدة، مقياس آخر للتأكد يقدم لكل قاعدة ممكنة مولدة. ويعتمد مقياس التأكد المقترح الذي استخدمناه هنا على دالة العضوية التقريبية:

$$\mu_X^I(x) = \frac{|X \cap I(x)|}{|I(x)|}$$

وللتوضيح سوف نقوم بتوليد قواعد القرار غير المحددة لنظام المعلومات الموضح بالجدول (1).

وجدنا أن المنطقة الموجبة لصفة القرار D هي:

$$POS_C(D) = \bigcup_{X \in U/D} \underline{Apr}(X) = \{x1, x3, x4, x6\}$$

$$\Rightarrow T - POS_C(D) = NEG_C(D) = \{x2, x5\}$$

وبعد حساب المختزل وجدنا بأن أحد المختزلات القوية هو المختزل الموضح في الجدول (3).عندها ستكون قواعد القرار غير المحددة على الشكل:

$$r : Condition(r) \xrightarrow{\mu_X^B(x)}{}^{S(r)} Decision(r)$$

حيث أن S(r) تمثل قوة القاعدة الممثلة لعدد الأمثلة الموجبة، و $\mu_X^I(x)$ تمثل درجة التأكد لكل قاعدة.بعد الحساب نجد أن: $\mu_D^I(x2) = \mu_{NotD}^I(x5) = 1/2$. وبالتالي تكون القواعد غير المحددة المولدة لنظام المعلومات بالجدول (1) هي:

$$r1: (A = a0) \wedge (C = c0) \xrightarrow{1}_{1/2} (D = d1)$$

$$r2: (A = a0) \wedge (C = c0) \xrightarrow{1}_{1/2} (D = d0)$$

7 طريقة $C_{5.0}$ والطريقة المقترحة:

طريقة$C_{5.0}$ هي خوارزمية استقراء قوية، اقترحتوطورت من قبل كوينلان [16, 17]. وهي خوارزمية محسنة عن$C_{4.5}$، التيهي خوارزمية محسنة عن ID3.تستخدم طريقة $C_{5.0}$ استراتيجية فرق تسد لإنتاج قواعد القرار، التي منها يمكن أن تشتق القواعد الواضحة. و$C_{5.0}$ هي امتداد لـ$C_{4.5}$ التي تفسر القيم غير المتاحة، ومجال القيم للصفات المستمرة، وتبني أشجار القرار وتقلمها، وتشتق القواعد من الأشجار المقلمة. في بناء شجرة القرار، يمكن أن نتعامل مع مجموعات التدريب التي تتلك سجلات لها قيم صفات مجهولة بتقييم الربح، أو نسبة الربح، لصفة معينة بأخذ بعين الاعتبار فقط تلك السجلات التي تكون قيم صفاتها متاحة. ويمكن أن نصنف السجلات التي تملك قيم صفات مجهولة بإزالة احتمالية النتائج الممكنة المختلفة.

بخلاف $C_{5.0}$، تولد طريقتنا المقترحة قواعد القرار المحددة وغير المحددة مباشرة، بدون توليد شجرة القرار، بناءً على مفاهيم المختزل، ومقياس قوة القاعدة S(r) ومقياس التأكد، الذي يعتمد على دالة العضوية التقريبية $\mu_X^I(x)$. وبناءً على ذلك، تعتبر طريقتنا المقترحة أسهل وأسرع من$C_{5.0}$، كونها لا تمر بمرحلة بناء أشجار القرار وتقليمها، التي منها تشتق قواعد القرار في $C_{5.0}$.

8 الخاتمة:

قدمنا في هذه المقالة تصميم إطار عمل مقترح لاكتشاف القواعد المحددة والقواعد غير المحددة من قواعد المعطيات اعتماداً على نظرية المجموعات التقريبية حيث تم في هذه المقالة بناء وتطوير عدة خوارزميات جديدة أهمها خوارزميتين هما خوارزمية توليد القواعد المحددة وخوارزمية توليد القواعد غير المحددة.

المراجع:

1. Frege, G., 1970- Grudgesetze der Arithmetik 2, in Geach and Black (Ed) Selections from Philosophical Writings of GotolobFrege, Blackweil, Oxford,.
2. Xiaohua Hu and Nick Cercone, 1995-" Learning in Relational Databases: A Rough Set Approach" In Computational Intelligence :An International Journal, Volume 11, Issue 2, May, pp. 323-338.
3. X.Hu, N. Cercone, J. Han, 1994- A Rough Set Approach for Knowledge Discovery in Databases, Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer_Verlag Press, W, Ziarko(ed), 90-99.
4. MertBal, 2013- Rough Sets Theory as Symbolic Data Mining Method: An Application on Complete Decision Table, An International Journal: Information Science Letters , 35-47.
5. ZdzislawPawlak, 1991- "Rough Sets: Theoretical Aspects of Reasoning about Data", Kluwer, Boston.
6. Pawlak, Z. and Skowron, A., 2007- Rudiments of Rough Sets, Information Sciences, 177,3-27.
7. Pawlak, Z. and Skowron, 2007- A. Rudiments of Rough Sets: Some Extensions, Information Sciences, 177, 28-40.
8. Pol and Sko, 1998- "Rough Sets in Knowledge Discovery 2 :Applications, Case and Software systems". Warsaw, February, Ch3,4 px39-46,px58-60.
9. Slowinski, R., 1992- Intelligent Decision Support: Handbook of Advances and Applications of the Rough Sets Theory, Kluwer Academic Publishers, Boston.
10. Yao, J. T.; Yao, Y. Y., 2002- "Induction of Classification Rules by Granular Computing". Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing (TSCTC'02). London, UK: Springer-Verlag. pp. 331–338.
11. Yao, Y. Y., 2006- Probabilistic Rough Set Approximations, Manuscrip.
12. Yao, J. T. and Herbert, J. T., 2007-Web-Based Support Systems Based on Rough Set Analysis, Manuscript.
13. Yao, Y.Y., 2010- Probabilistic Rough Set Approximations, International Journal of Approximations Reasoning, 49, 255-271.
14. Ziarko, W., 2008- Probabilistic Approach to Rough Sets, International Journal of Approximate Reasoning, 49, 272-284.

15. Max Kuhn, 2013- Classification Using C5.0, Pfizer Global R&D.
16. Shuonan Hou and Et al, 2014- Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics, International Journal of Control and Automation, vol. 7, No. 1, pp. 181-190.
17. RutvijaPandya and Jayati Pandya, 2015- C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, International Journal of Computer Applications, Volume 117, No. 16.