

TD : Induction supervisée de concepts : l'espace des versions

Soient des exemples correspondants à des jours possibles pour faire du sport et décrits par les attributs :

- *Ciel* (avec comme valeurs possibles : *soleil, nuages, pluie*)
- *AirTemp* (valeurs possibles : *chaud, froid*)
- *Humidité* : (valeurs possibles : *normale, élevée*)
- *Vent* (valeurs possibles : *fort, faible*)
- *Eau* (valeurs possibles : *chaude, fraîche*)
- *Prévision* (valeurs possibles : *égale, change*)

Soit l'espace des hypothèses décrites par des conjonctions de contraintes sur les attributs. Les contraintes peuvent prendre la forme

- "?" : toute valeur est possible
- "Ø" : aucune valeur n'est acceptable
- une valeur spécifique (e.g. Eau = chaude).

On cherche à trouver l'ensemble des hypothèses compatibles avec les données d'apprentissage (l'espace des versions), c'est-à-dire telles qu'elles couvrent les exemples positifs et excluent les exemples négatifs.

Soit la séquence d'exemples suivante :

<i>Exemple</i>	<i>Ciel</i>	<i>AirTemp</i>	<i>Humidité</i>	<i>Vent</i>	<i>Eau</i>	<i>Prévision</i>	<i>Classe</i>
1	Soleil	Chaud	Normale	Fort	Chaude	Egale	+
2	Soleil	Chaud	Elevée	Fort	Chaude	Egale	+
3	Pluie	Froid	Elevée	Fort	Chaude	Change	-
4	Soleil	Chaud	Elevée	Fort	Fraîche	Change	+

1. Expliquez pourquoi la taille de l'espace des hypothèses est de 973. Quel serait le nombre d'exemples possibles et d'hypothèses possibles si l'on ajoutait l'attribut *Courant_d_Eau* avec les valeurs *faible, modéré, fort* ? Plus généralement comment croissent le nombre d'exemples possibles et d'hypothèses quand on ajoute un attribut *A* pouvant prendre une valeur parmi *k* ?

2. Spécifiez le S-set et le G-set après la prise en compte de chaque exemple, par l'algorithme d'élimination des candidats de l'Espace des Versions.

3. Spécifiez le S-set et le G-set après la prise en compte de chaque exemple de la séquence prise *dans l'ordre inverse*, par l'algorithme d'élimination des candidats de l'Espace des Versions. Bien que l'espace des versions final soit le même (pourquoi ?), les ensemble S et G vont bien sûr dépendre de cet ordre. Pouvez-vous en tirer des idées sur l'optimisation de la séquence d'apprentissage en vue de minimiser la taille du S-set et du G-set ?

4. Supposons maintenant que l'on considère un nouvel espace d'hypothèses \mathcal{H}^1 consistant dans les hypothèses disjonctives à deux termes. Par exemple : $\langle ?, \text{froid, élevée}, ?, ?, ? \rangle \vee \langle \text{soleil}, ?, \text{élevée}, ?, ?, \text{égale} \rangle$

Calculez alors le S-set et le G-set pour la séquence d'exemples donnée plus haut.

On rappelle ci-dessous l'**algorithme d'élimination des candidats** :

Initialiser S et G par les ensembles des généralisations les plus spécifiques et les plus générales cohérentes avec le premier exemple positif

Pour chaque exemple suivant *i* :

Début

Si *i* est un exemple négatif **alors**

- ne retenir dans S que les généralisations ne couvrant pas *i*
- rendre plus spécifiques les généralisations de G couvrant *i* juste assez pour qu'elles ne couvrent plus *i*, et seulement de manière que chacune reste plus générale que certaines généralisations de S
- Retirer de G tout élément plus spécifique que d'autres éléments de G

fin alors

sinon, si *i* est un exemple positif **alors**

- ne retenir dans G que les généralisations couvrant *i*
- généraliser les éléments de S ne couvrant pas *i* juste assez pour leur permettre de couvrir *i*, et seulement de telle manière que chacune reste plus spécifique que certaines généralisations de G
- retirer de S tout élément plus général que d'autres éléments de S

fin si

fin pour chaque

5. Supposons que l'algorithme d'apprentissage par élimination des candidats effectue un *apprentissage actif*, c'est-à-dire puisse choisir soit l'ordre des exemples à prendre en compte parmi un ensemble d'exemples donnés, soit même soit capable de poser des questions comme au jeu de Mastermind.

Quelle serait, selon vous, une bonne stratégie de choix des exemples pour accélérer la convergence de l'apprentissage ?

6. On s'intéresse maintenant à l'induction de grammaires régulières qui sont représentables par des automates finis. Pouvez-vous définir une relation d'ordre partiel de généralité entre automates ?

