Metodologia de caracterització conceptual per condicionaments successius. Una aplicació a sistemes mediambientals

Alejandra Alicia Pérez Bonilla

Data de lectura:
12 / 01 / 2010

Directora
Karina Gibert

Programa de doctorat:
Doctorat en aplicacions tècniques i informàtiques de l'estadística, la investigació operativa i l'optimització

Universitat:
Universitat Politècnica de Catalunya

En classificació automàtica, es busquen perfils subjacents a l'estructura d'un domini que ajudi a comprendre'l i que permeti una millor presa de decisions. Tanmateix, comprendre el significat de les classes resulta fonamental. D'altra banda, la validació d'un cluster segueix sent un problema obert, doncs no s'ha trobat encara un criteri objectiu per a determinar la qualitat d'un conjunt de classes en el context del clustering (Hand 1996), i que s'aplica en situacions on no hi ha un bon coneixement de l'estructura del domini. Volle al 1985, fa tota una dissertació, il·lustrant que, el concepte de validesa no és absolut, sinó relatiu a les condicions del context i a la utilitat de les classes descobertes. Tot i que aquest és un extrem poc objectivable, la interpretació constitueix una fase fonamental del procés i segueix sent, encara avui, un dels criteris més utilitzats en la pràctica per a validar el cluster. Per aquesta raó, la validació queda directament lligada a l'existència d'una interpretació clara per al clustering o partició.

Així, es fa necessari introduir eines per a assistir a l'usuari en les tasques d'interpretació d'una partició sobre un conjunt d'objectes, amb la finalitat d'establir el significat de les classes resultants. Si les classes obtingudes no tenen sentit per a l'expert(s), els resultats de la classificació no són considerats vàlids, ni tampoc es podran utilitzar, ni donaran suport a cap decisió posterior. Totes les tècniques i algorismes de validació existents a la literatura, van orientats al vessant estructural de la partició; però disposar de classes ben formades estructuralment, no oferix la garantia que un expert sigui capaç d'associar cadascun d'aquests grups a una entitat semàntica. Aquesta tesi pretén contribuir a la millora d'aquest procés, fonamental per a comprendre el significat de les classes obtingudes i donar suport efectiu a la posterior presa de decisions.

L'alternativa que sembla més prometedora, per resoldre aquestes limitacions, és alleugerir l'expert de l'anàlisi en brut de les classes obtingudes, mitjançant el desenvolupament de tècniques que, a partir de l'evidència empírica, identifiquin les variables més rellevants i formulin conceptes relatius a les particularitats de cada classe i s'expressin en una representació conceptual generable automàticament i directament comprensible per a l'expert.

La tesi adopta un enfoc de Knowledge Dicovery from Data (KDD), segons el qual la fase de post-procés dels resultats per generar coneixement és gairebé tan important com l'anàlisi en sí mateix. Potser, per la seva naturalesa més semàntica, la generació automàtica d'interpretacions d'una classificació no s'ha tractat formalment des de l'àmbit estadístic, encara que resoldre'l és fonamental.

En aquesta tesi es proposa una solució aproximada al problema de construir un sistema de conceptes $A_{P_\xi} = \{A_1, A_2, ..., A_\xi\}$ que descriuen les classes de tal manera que, donada una partició en ξ classes, $P_\xi = \{C_1, C_2, ..., C_\xi\}$, sobre un conjunt d'objectes $i \in I$, obtinguda per classificació jeràrquica:

- $A, A' \in A_{P_\xi} \Rightarrow A \neq A'$
- $\forall i \in I, A_C(i) = \text{cert, si } C = C(i, P_\xi), A_C \in A_{P_\xi}$
- $\forall i \in I, A_C(i) = \text{fals, si } C \neq C(i, P_\xi), A_C \in A_{P_\xi}$

Tenint en compte que existirà certa incertesa en el model, es proposa tractar amb regles més genèriques de la forma $r : A_C(i) \xrightarrow{p} C,\ \text{on } p \in [0,1]$ on $p \in [0,1]$ és la probabilitat amb que es compleix **r**. D'aquesta manera les regles incorporen incertesa sota una aproximació probabilística.

La metodologia que es proposa tracta d'aproximar a un model formal el procés natural que segueix un expert en la seva fase d'interpretació de resultats, realitzant una aproximació iterativa basada en la classificació jeràrquica subjacent. La proposta que es presenta:

- Aporta una sistematització al procés d'interpretació de classes procedents d'un cluster jeràrquic i suposa un avenç significatiu respecte a l'estat actual en que la interpretació es realitza de forma manual i més o menys artesanal.
- Així mateix, contribueix a sistematitzar i objectivar els mecanismes d'interpretació que usen els experts humans.
- Els resultats que genera la metodologia permeten que l'expert pugui comprendre més fàcilment les característiques principals de la classificació obtinguda ja que genera coneixement explícit directament a partir de les classes.

Si bé la metodologia que es proposa és general, s'ha centrat l'aplicació a Estacions depuradores d'aigües residuals (EDAR) per ser aquest un dels dominis on les aproximacions clàssiques funcionen pitjor i perquè s'enquadren en una de les línies marc d'investigació que es desenvolupa en el grup.

Des d'un punt de vista teòric, l'interès d'aquesta tesi ha estat en presentar una proposta metodològica híbrida que combini eines i tècniques d'Estadística i d'Intel·ligència Artificial en forma cooperativa, seguint un enfocament transversal i multidiciplinar combinant elements d'inducció de conceptes, lògica proposicional i teoria de probabilitat. És així com, aquesta tesi, representa una contribució a la concepció genèrica de sistema integral de KDD, segons el marc general introduit per Fayyad. També contribueix a objectivar els procediments de validació de classes, per la relació entre la clara interpretació dels clusters i la utilitat d'una classificació (que actualment també es fa servir com criteri de validació); avaluar el clustering requereix un mecanisme a posteriori de comprensió del significat de les classes.

La metodologia de Caracterització Conceptual per Condicionaments Successius (CCCS) aprofita l'estructura jeràrquica de la classificació objectiu per a induir conceptes iterant sobre les divisions binàries que indica el dendrograma, seleccionant a cada iteració les variables que millor distingeixen cada classe i incorporant herència en les iteracions successives. A partir de les variables que descriuen els objectes pertanyents a cert domini i l'anàlisi automàtica de les seves distribucions condicionades, es troben les particularitats de cada classe i es proposa una solució aproximada al problema inicial de la forma $A_{P_\xi} = \{C : A_C \forall C \in P_\xi\}$, on A_C són conceptes que permeten entendre i distingir les classes, a partir d'una construcció jeràrquica d'un conjunt de regles $R(P_\xi) = \left\{ r \text{ tq } r : A \xrightarrow{p(r)} C, \forall C \in P_\xi \right\}$.

En el problema que s'intenta resoldre, no és tan important el poder discriminatori i predictiu de les variables que intervindran en la solució, com la seva capacitat explicativa, el que té conseqüències directes sobre els criteris de selecció de variables de la proposta.