Adatbányászati algoritmusok

2002/2003. tanév II. félév

VIZSGA FELADATOK – 2003. MÁJUS 27.

(1) Az APRIORI algoritmus melyik hibájára ad megoldást a DIC algoritmus (**2 pont**)? Írja le a DIC működését (**3 pont**)! Mutasson egy konkrét példát arra, amikor a DIC algoritmus jobb az APRIORI-nál (**1 pont**)!

(2) Előfordulhat-e, hogy az egyszerű mintavételező algoritmus lefutása után, biztosak lehetünk abban, hogy minden gyakori termékhalmazt megtaláltunk, ha a minta kisebb a teljes adatbázisnál és a mintában nem gyakori minden termékhalmaz? Válaszát indolkolja (**3 pont**)!

(3) Vázolja a Toivonen algoritmus működését (**5 pont**)! Mutasson példát arra, amikor az algoritmus az adatbázis egyszeri végigolvasásával megtalálja az összes jelöltet és mutasson példát arra is, amikor ehhez két végigolvasás szükséges (**1 pont**). Írja le, hogy kik lesznek a jelöltek ebben az esetben (**1 pont**). Hogyan alkalmazhatjuk Toivonen algoritmusának ötletét az on-line asszociációs szabálybányászatban? (**4 pont**)?

(4) Hogyan definiáljuk a Duquenne–Guigues-bázist (**2 pont**)? Legyen az adatbázisunkban 3 kosár: $\{A, C\}, \{B\}, \{A, B, D\}$. Adja meg a zárt és a pszeudó-zárt termékhalmazokat (**3 pont**)! Igaz-e, hogy a zárt asszociációs szabálygráf egy tetszőleges lefedő élhalmazához tartozó asszociációs szabályokból az összes közelítő asszociációs szabály levezethető? Válaszát indokolja (**4 pont**)!

(5) Hogyan definiáljuk és mi célt szolgálnak egy sorozat összefüggő részsorozatai (**2 pont**)? Írja le a GSP jelöltgenerálásának menetét (**2 pont**). Hogyan általánosíthatjuk az epizódok fogalmát (**2 pont**)?

(6) Vázolja mi történik a lenyomat alapú hasonlóságkeresés 3 lépésében (**2 pont**)? Hogyan állítjuk elő a min-hash lenyomatokat (**5 pont**)? Mennyire jó ez az eljárás, mit tudunk mondani a helyességéről (**3 pont**)?

(7) Hogyan mérhetjük egy közelítő függőség hibáját (**3 pont**)? Hogyan definiáltuk a TANE algoritmusban a partíciók szorzatát és mikor hívunk egy attribútumhalmazt kulcsnak (**3 pont**)? Hogyan dönti el a TANE algoritmus, hogy egy függőség minimális-e (**6 pont**)?

(8) Milyen elvárásaink vannak egy jó klaszterező algoritmussal szemben (**4 pont**)? Írja le a k-közép működését. Milyen hibái vannak az algoritmusnak (**5 pont**)? Hogyan definálhatjuk egy klaszter átmérőjét, illetve a klaszterek közötti távolságot (**2 pont**)? Mi jellemző a sűrűség alapú klaszterező eljárásokra (**2 pont**)?