Szépe Tamás

Szegedi Tudományegyetem

Konzulens: Dr. Kocsor András
tudományos főmunkatárs

AUTOMATIKUS KLASZTEREZÉSEN ALAPULÓ JELLEMZŐ KIVÁLASZTÁS

Számos gépi tanuláshoz használt algoritmus készült, melyekkel valamilyen megfigyelési sorozatot lehet csoportokra bontani, klaszterezni. Ezen algoritmusok hatékonyságát nagymértékben befolyásolja a mintahalmaz, amin hipotézist állítanak elő. A számítási bonyolultság, vagy a kis elemszámú tanulóhalmaz megnehezíti a rendelkezésre álló minták összes paraméterének felhasználását. A mintasűrűség szinten tartása a dimenzió lineáris növelése mellett csak a megfigyelések exponenciális növelésével valósítható meg. Ezt a jelenséget nevezzük a „dimenzionalitás átkának”.

Ha a magas dimenzió miatt nincs meg a kellő mintasűrűség, akkor az algoritmusok pontatlanná, vagy lassúvá válnak. Ennek megoldására nyújt lehetőséget a jellemző kiválasztása (*feature selection*) eljárás. Elméleti úton belátható, hogy felügyelt tanulás esetén optimális megoldás csak kimerítő kereséssel adható, ezért gyakorlatban elfogadható közelítő eredmény is. A legnépszerűbb megközelítések a problémára a mohó hegymászó algoritmus implementációi.

Ebben a dolgozatban megismerhetünk egy olyan nemlineáris algoritmust, mely folyamatos közelítést alkalmazva képes csökkenteni a mintahalmaz dimenziószámát, kiválasztva a leglényegesebb tulajdonságokat. Az eljárás alapjaként az X-Means algoritmus többszöri futtatása szolgál. Optimalizációs kritériumként a benne felhasznált Bayes információs kritérium (BIC), valamint felügyelt esetre optimalizált saját kritérium szerepel. Az eljárás kellően univerzális ahhoz, hogy használható legyen felügyelt, és felügyelet nélküli tanuló algoritmusokhoz egyaránt, illetve alkalmazható hierarchikus hipotézisek kereséséhez is. Az eljárások létjogosultságát szimulációs eredményekkel igazolom.