Marta Zalewska
Akademia Medyczna w Warszawie, Zakład Profilaktyki Zagrożeń Środowiskowych
Antoni Grzanka
Politechnika Warszawska, Instytut Systemów Elektronicznych
Wojciech Niemiro
Uniwersytet Mikołaja Kopernika w Toruniu, Wydział Matematyki i Informatyki
Bolesław Samoliński
Akademia Medyczna w Warszawie, Zakład Profilaktyki Zagrożeń Środowiskowych

Identyfikacja nietypowych podzbiorów danych z zastosowaniem w badaniach medycznych

W wielu dziedzinach nauki, m.in. w medycynie, socjologii, ekonomii, pojawia się potrzeba identyfikacji „odstających" (outlying) podzbiorów danych. Problem jest szczególnie ważny, jeśli mamy duże zbiory danych. O ile literatura na temat identyfikacji odstających obserwacji jest obszerna, to wykrywanie nietypowych *podzbiorów* obserwacji jest niedostatecznie zbadane. Gotowe pakiety statystyczne nie dostarczają odpowiednich narzędzi do rozwiązania tego problemu. Proponujemy metodę opartą na oszacowaniu stopnia oddzielenia (separacji) wyróżnionego podzbioru od reszty danych. Zdefiniowana przez nas miara separacji jest związana z estymatorem typu „jackknife" prawdopodobieństwa błędnej klasyfikacji dla kwadratowej funkcji dyskryminacyjnej. Nasz algorytm należy do metod „komputerowo intensywnych" i wykorzystuje wielokrotne repróbkowanie (resampling) do obliczenia poziomu istotności odpowiedniego testu statystycznego. Metoda zostanie zilustrowana na przykładzie identyfikacji odstających podzbiorów ankiet z badań alergologicznych.