Metody grupowania danych i ich wybrane modyfikacje dedykowane analizie danych eksperymentalnych

Autor: Klaudia Drab
Promotor pracy: dr hab. Michał Daszykowski, prof. UŚ

Zaawansowana aparatura badawcza umożliwiła badanie materiałów różnorodnego pochodzenia. Dlatego znalazła ona zastosowanie w wielu dziedzinach nauki, gdzie stanowi podstawowe narzędzie w ocenie fizykochemicznych właściwości próbek. Jednak kompleksowa charakterystyka próbki pociąga za sobą pozyskiwanie danych o złożonej strukturze. Opisując każdą analizowaną próbkę za pomocą od kilku do kilku tysięcy zmiennych otrzymuje się tzw. dane wielowymiarowe, co pociąga za sobą potrzebę zastosowania metod matematycznych, pozwalających na analizę i interpretację wyników oraz formułowanie wniosków. W tym celu korzysta się z metod chemometrycznych, w skład których wchodzą metody wstępnego przygotowania danych do analizy, metody eksploracyjne oraz metody modelowania danych. Szczególnie interesujące są metody eksploracyjne pozwalające na wgląd w ukrytą strukturę analizowanych danych oraz ujawnienie zależności pomiędzy próbkami i/lub parametrami. Jednym z wariantów metod eksploracyjnych są metody grupowania danych, które są szczególnie przydatne w kontekście wyodrębniania grup podobnych obiektów. Podobieństwo analizowanych próbek, oceniane jest na podstawie ich odległości w przestrzeni eksperymentalnej (te które znajdują się blisko siebie wykazują podobne właściwości fizykochemiczne). W celu określenia podobieństwa pomiędzy próbkami wykorzystuje się tzw. miary podobieństwa, które są matematyczną interpretacją odległości pomiędzy nimi.

Wzrastająca kompleksowość danych pociąga za sobą potrzebę modyfikacji i rozwoju nowych metod eksploracyjnych oraz miar odległości. W związku z czym w niniejszej pracy doktorskiej skupiono się na modyfikacji algorytmu DBSCAN w celu eliminacji problemu błędnego przypisania obiektów brzegowych do odpowiednich grup w przypadku grup sąsiadujących ze sobą w przestrzeni eksperymentalnej. Modyfikacja algorytmu polegała na zmianie sposobu przetwarzania obiektów oraz przypisaniu obiektów brzegowych do grup na podstawie odległości euklidesowej pomiędzy obiektami brzegowymi, a środkami wyodrębnionych grup obiektów. Następnie skupiono się na rozwinięciu koncepcji nowej miary odległości (s_{ij}) pozwalającej porównywać ze sobą dwuwymiarowe chromatograficzne odciski palca, w których występuje problem koelucji substancji i przesunięć pików w czasie. W ostatniej części pracy rozważano problem niepewności pomiarowej towarzyszącej danym eksperymentalnym. Dotychczas, błąd pomiarowy był pomijanym elementem w trakcie analizy danych. Aktualnie rozwój algorytmów uwzględniających niepewności pomiarowe uzyskiwanych danych stanowi nowy trend w pracach naukowych. Korzystając z osiągnieć zaprezentowanych w [1], zaproponowano uwzględnienie niepewności pomiarowych modelowanych dla każdego obiektu np. w algorytmie DBSCAN, poprawiając efektywność metody.

[1] M. Kumar, N. R. Patel, Clustering data with measurement errors, Comput. Stat. Data Anal. 51 (2007) 6084–6101.