SILESIAN UNIVERSITY OF TECHNOLOGY

FACULTY OF AUTOMATIC CONTROL, ELECTRONICS AND COMPUTER SCIENCE

PHD THESIS

DATA CLUSTERING WITH MIXTURES OF MULTIDIMENSIONAL DISTRIBUTIONS

mgr Mateusz Kania

promoter: prof. zw dr hab inż Andrzej Polański

Grupowanie danych z wykorzystaniem mieszanin wielowymiarowych rozkładów

Streszczenie

Nienadzorowane grupowanie jest ważnym obszarem w analizie danych, uczeniu maszynowym i sztucznej inteligencji. Celem tego projektu doktoranckiego było opracowanie, implementacja i porównanie algorytmów nienadzorowanego klastrowania opartych na modelu i na odległości oraz ocena ich wydajności na różnych zbiorach danych przy użyciu różnych metryk. Badanie miało na celu rozwiązanie wyzwania, przed którym stoją naukowcy zajmujący się danymi przy wyborze odpowiedniego algorytmu nienadzorowanego klastrowania, biorąc pod uwagę wiele dostępnych metod, którym często towarzyszą implementacje oprogramowania.

W pracy zaimplementowano dwa algorytmy oparte na wielowymiarowych modelach mieszanin, Gaussian Mixture EM i Multinomial Mixture EM, i porównano je z czterema algorytmami opartymi na odległości, aglomeracyjnym grupowaniem hierarchicznym, k-means, k-medoids i rozmytym c-means. Algorytmy zostały zastosowane zarówno do symulowanych, jak i rzeczywistych zbiorów danych, a kilka metryk zostało wykorzystanych do ilościowej oceny wyników grupowania, w tym skorygowany indeks Randa, współczynnik prostego dopasowania, ważony indeks Jaccarda, zrównoważona dokładność oraz metryki oparte na sprzężonym rozkładzie beta-binomialnym. Wyniki pokazały, że algorytmy oparte na modelu, w szczególności Gaussian Mixture EM i Multinomial Mixture EM, przewyższają w wielu przypadkach algorytmy oparte na odległości.

Wkładem badania w dziedzinę analizy danych i uczenia maszynowego było zapewnienie wglądu w rozwój dokładniejszych i skuteczniejszych metod nienadzorowanego grupowania. Badanie wykazało, że algorytmy oparte na modelach są potężnymi narzędziami w metodach nienadzorowanego grupowania i są wysoce konkurencyjne w porównaniu z algorytmami opartymi na odległości. Ponadto, algorytmy zostały zaimplementowane w języku R i udostępnione na platformie GitHub.

W pracy przeprowadzono obszerne badanie symulacyjne z wykorzystaniem tysięcy wielomianowych mieszanek gaussowskich i wielomianowych w celu przetestowania wydajności algorytmów. Dodatkowo, porównano również kuratorski zestaw rzeczywistych zbiorów danych z różnych publicznie dostępnych źródeł, w tym danych genomicznych/medycznych. Na podstawie tych zbiorów danych przygotowano setki różnych komponentów, przy czym każda kombinacja grup występuje tylko raz w tym samym zbiorze, co pozwoliło na kontrolowane porównanie algorytmów przy różnej liczbie parametrów, wymiarów i klastrów.

Podsumowując, wyniki badania wykazały, że algorytmy oparte na modelach, w szczególności Gaussian Mixture EM i Multinomial Mixture EM, w wielu przypadkach przewyższają algorytmy oparte na odległości. Badanie przyczynia się do postępu w dziedzinie analizy danych i uczenia maszynowego poprzez informowanie o rozwoju bardziej dokładnych i skutecznych metod nienadzorowanego grupowania. Opracowane algorytmy są dostępne w języku R i mogą być wykorzystywane w różnych aplikacjach, przyczyniając się do rozwiązywania wyzwań w różnych dziedzinach nauki.