Streszczenie

Szybki rozwój zaawansowanych metod chemii analitycznej otwiera nowe możliwości analizy próbek mikrośladów (np. fragmentów polimerów, lakierów samochodowych, okruchów szkła) dla celów sądowych. Udoskonalenie technik analitycznych pozwala na rejestrację tysięcy parametrów opisujących badane próbki w stosunkowo krótkim czasie. Wraz z postępem w tej dziedzinie, koniecznością staje się również rozwój i dostosowanie technik interpretacji tak ogromnej ilości danych, szczególnie, gdy wnioskowanie dotyczy nauk sądowych.

Analiza i interpretacja dużych zbiorów danych najczęściej wymaga wstępnego zastosowania technik redukcji ich wymiarowości za pomocą metod chemometrycznych. Ich celem jest uwypuklenie ukrytej struktury danych oraz wydobycie jak najcenniejszej informacji niesionej przez dane analityczne, a dotyczącej podobieństwa analizowanych obiektów w postaci jak najmniejszej liczby nowych zmiennych. Zadania te są z powodzeniem realizowane przez znakomitą liczbę metod chemometrycznych, niemniej jednak ich bezpośrednia aplikacja w naukach sądowych nie jest możliwa i wymaga wprowadzenia pewnych modyfikacji, by móc stanowić podstawę wnioskowania w procesie sądowym. Problem ten sprowadza się głównie do trudności w uwzględnieniu częstości występowania określonych cech w całej populacji generalnej analizowanych materiałów, korelacji między zmiennymi, jak i możliwych źródeł zmienności. Brak tych elementów powoduje, że zastosowaniu metod chemometrycznych musi towarzyszyć ocena wartości dowodowej przeprowadzona z zastosowaniem metodologii opartej o testy ilorazu wiarygodności (ang. likelihood ratio, LR).

Podejście to pozwala na uwzględnienie wszystkich czynników niezbędnych z punktu widzenia wymiaru sprawiedliwości, w tym częstości występowania określonych cech w całej populacji generalnej, możliwych źródeł zmienności oraz korelacji między zmiennymi. Mimo szeregu zalet, jedną z wad modeli LR jest trudność ich konstrukcji dla danych o dużej

wymiarowości, gdy liczba zmiennych zdecydowanie przekracza liczbę próbek, które opisują. Przykładem tego rodzaju danych mogą być widma lub chromatogramy uzyskane w wyniku zastosowania popularnych spektroskopowych lub chromatograficznych technik analitycznych. Problem wielowymiarowości związany jest głównie z trudnościami w rzetelnym oszacowaniu parametrów populacyjnych takich jak średnie, wariancje lub kowariancje. Dlatego też celem badań opisanych w rozprawie było opracowanie metodologii pozwalającej na ocenę wartości dowodowej wielowymiarowych danych fizykochemicznych z wykorzystaniem technik chemometrycznych uwidaczniających strukturę danych i eksponujących najistotniejsze cechy w postaci niewielkiej liczby zmiennych stanowiących bazę do konstrukcji modeli ilorazu wiarygodności, będących standardem w ocenie wartości dowodowej danych w naukach sądowych.

W badaniach podjęto się rozwiązania tzw. problemu porównawczego, w którym formułowane są następujące hipotezy:

H_1: porównywane próbki pochodzą z tego samego źródła (np. samochodu),

H_2: porównywane próbki nie pochodzą z tego samego źródła.

Modele LR zostały zaprojektowane dla bazy danych próbek polimerów zbudowanych w taki sposób, aby jak najwierniej odzwierciedlały rzeczywiste materiały zabezpieczane na miejscu np. wypadku drogowego. Dlatego też część próbek pochodziła z plastikowych elementów nadwozia samochodów (np. zderzaków), a bazę uzupełniała grupa plastikowych pojemników, stanowiących opakowania produktów codziennego użytku (np. kosmetyków), które mogą także stanowić źródło polimerów na miejscu np. wypadku drogowego. W badaniach skupiono się na jednym z najpopularniejszych typów polimerów, jakim jest polipropylen, nie tylko ze względu na jego szerokie zastosowania w przemyśle motoryzacyjnym, ale również ze względu na prostą strukturę uwidaczniającą się w przejrzystych i nieskomplikowanych widmach i chromatogramach. Wbrew pozorom taki wybór materiału stanowił wyzwanie w przypadku, gdy jest on podstawą problemu porównawczego. Jedna z baz danych zawierała informacje o 27 próbkach polipropylenowych analizowanych z wykorzystaniem fourierowskiej spektrometrii w podczerwieni (ang. Fourier transform infrared spectrometry, FTIR), natomiast druga dotyczyła 25 próbek analizowanych z wykorzystaniem pirolitycznej chromatografii gazowej sprzężonej ze spektrometrem mas (ang. pyrolysis gas chromatography mass spectrometry, Py-GC-MS).

Głównym założeniem pracy było zaproponowanie takich metod oceny wartości dowodowej w problemie porównawczym widm i chromatogramów polipropylenu, by sformalizować i zobiektywizować stosowaną do tej pory wizualną, a więc subiektywną, ocenę ich podobieństwa. W tym celu skonstruowano kilka hybrydowych modeli LR łączących w sobie zalety technik chemometrycznych oraz podejścia opartego o teorię ilorazu wiarygodności. W badaniach zweryfikowano użyteczność analizy głównych składowych (ang. principal component analysis, PCA), dyskretnej transformacji falkowej (ang. discrete wavelet transform, DWT), liniowej analizy dyskryminacyjnej (ang. linear discriminant analysis, LDA) oraz reprezentacji odległościowej (ang. distance representation, DR) do uzyskania zredukowanej liczby zmiennych w sposób wyczerpujący opisujących cechy analizowanych próbek.

W pierwszym modelu zastosowano dyskretną transformację falkową jako metodę redukcji wymiarowości przestrzeni cech. Jej niekwestionowaną zaletą jest możliwość generowania uproszczonej formy widma o mniejszej liczbie danych, ale jednocześnie zachowującej najistotniejsze cechy z punktu widzenia ich interpretacji chemicznej. Zignorowanie współczynników DWT o niskiej amplitudzie pozwoliło na efektywną redukcję wymiarowości danych, a analiza relacji pomiędzy zmiennością między- i wewnątrz-obiektową wspomogła wybór najistotniejszych zmiennych. Rezultaty wykazały, że ostatecznie wybrane zmienne grupują się w trzy zbiory, każdy odnoszący się do odrębnego fragmentu widma. Wśród zaproponowanych modeli jedno-, dwu- i trójwymiarowych, te ostatnie charakteryzowały się najniższym wskaźnikiem błędów fałszywie pozytywnych oraz fałszywie negatywnych. Ponadto obserwacje te zostały potwierdzone stosując empiryczną entropię krzyżową jako metodę uwzględniającą siłę wsparcia każdej z rozpatrywanych hipotez.

W drugim modelu klasyczna reprezentacji zmiennych została przekształcona w reprezentację odległościową, w której widma przedstawiono w postaci ich odległości od zestawu widm referencyjnych uwypuklających strukturę ich wzajemnego podobieństwa. Dane w reprezentacji odległościowej zdefiniowanej przez odległość Manhattan, Euklidesa, kwadrat odległości Euklidesa, Chebysheva oraz bazującą na współczynniku korelacji zostały poddane dodatkowo analizie LDA w celu zoptymalizowania relacji między zmiennością między- i wewnątrz-obiektową. Mimo iż metoda ta służy do celów klasyfikacyjnych, jej zdolność do maksymalnej separacji próbek wykorzystano dzięki potraktowaniu każdej próbki jako osobnej klasy. Dla tak przygotowanych danych zaproponowano tzw. naiwne modele LR, w których liczbę zmiennych wytypowano na podstawie algorytmu opartego na kryterium Bayesowskim

(ang. Bayesian Information Criterion, BIC). Uzyskane rezultaty podkreśliły przydatność zaproponowanej metodologii, przy czym zwróciły uwagę również na ograniczenia w stosowaniu empirycznej entropii krzyżowej do oceny efektywności modeli zbudowanych dla baz danych o niewielkiej liczbie próbek.

Modele LR zaproponowane dla danych uzyskanych z Py-GC-MS zostały zaczerpnięte z metodologii rozwiązywania problemu klasyfikacyjnego w naukach sądowych za pomocą testu ilorazu wiarygodności. Dane zostały poddane analizie PCA przeprowadzonej na średnich dla próbek, by niepotrzebnie nie zwiększać zmienności wewnątrz-obiektowej. W kolejnym etapie dane zostały poddane analizie LDA oraz przedstawione w reprezentacji odległościowej, gdzie odległość zdefiniowana była między każdymi dwoma chromatogramami poddanymi obróbce chemometrycznej. Modele LR skonstruowano w dwojaki sposób – jeden uwzględniał klasyczne podejście LR, natomiast w drugim wartość LR uzyskiwano ze stosunku wartości prawdopodobieństw *a posteriori* obliczonych na podstawie modelu regresji logistycznej. Uzyskane rezultaty potwierdziły użyteczność zaproponowanej metodologii, zarówno pod kątem satysfakcjonujących poziomów błędnych wskazań, jak i redukcji utraty informacji obserwowanej na wykresach empirycznej entropii krzyżowej.

Ponadto opracowano modele hybrydowe pozwalające na ocenę łącznej wartości dowodowej danych FTIR oraz Py-GC-MS tzw. analizę dowodu łączonego. Problem ten nabiera na znaczeniu szczególnie w naukach sądowych, gdzie uznaną procedurą jest analiza próbek z wykorzystaniem dwóch odrębnych metod analitycznych o odmiennych podstawach fizykochemicznych. Model ten bazował na wieloblokowej analizie PCA (ang. mulitblock PCA), której wyniki zostały poddane analizie LDA w celu uzyskania optymalnej separacji próbek. Ostatecznie wartość dowodową oceniono stosując test LR, którego efektywność potwierdziła fakt, iż uwzględnienie wyników pochodzących z więcej niż jednej metody pozwala na uzyskanie bardziej rzetelnych rezultatów stanowiących mocniejszą podstawę wnioskowania sądowego.

Pionierskim elementem przeprowadzonych badań było połączenie zalet metod chemometrycznych, szczególnie w odniesieniu do redukcji wymiarowości przestrzeni cech, z modelami ilorazu wiarygodności, które stanowią szeroko akceptowaną metodę oceny wartości dowodowej. Utworzenie hybrydowych modeli LR pozwoliło nie tylko na obiektywizację stosowanej do tej pory metody wizualnego porównywania widm lub chromatogramów w celu określenia ich podobieństwa, ale również ocenę tego podobieństwa

w sposób ilościowy wyrażony w postaci wartości LR informujących o sile wsparcia dla rozpatrywanych hipotez. Istotnym elementem badań było odniesienie się w zaproponowanych modelach nie tylko do zmiennych uzyskanych w toku przeprowadzonych zabiegów chemometrycznych, niejednokrotnie pozbawionych interpretacji chemicznej, ale do informacji chemicznej niesionej przez widma, co dodatkowo wpływa korzystnie na wiarygodność proponowanych modeli hybrydowych. Warto zwrócić uwagę, iż w badaniach nietypowo wykorzystano zalety liniowej analizy dyskryminacyjnej do maksymalnej separacji próbek w problemie porównawczym, mimo iż technika ta należy do metod klasyfikacyjnych. Ponadto w modelach skonstruowanych dla danych uzyskanych z techniki Py-GC-MS zaproponowano pionierskie rozwiązanie problemu porównawczego korzystając z klasycznej koncepcji problemu klasyfikacyjnego w naukach sądowych, co pozwoliło na zmniejszenie stopnia skomplikowania i złożoności obliczeniowej modeli.

Przedstawione wyniki badań jednoznacznie wskazują, iż hybrydowe modele LR łączące metody chemometryczne z modelami ilorazu wiarygodności pozwalają na rozwiązanie problemu porównawczego baz danych próbek polipropylenu opisanych za pomocą widm FTIR i chromatogramów Py-GC-MS mimo niewielkiej liczby zgromadzonych próbek, co również jest elementem nowości opisanych badań. Problem ten staje się szczególnie istotny ze względu na coraz większą różnorodność materiału dowodowego, co pociąga za sobą konieczność budowy odpowiednich baz danych. Dlatego też zastosowanie metodologii niewymagających dużej liczby próbek do uzyskania rzetelnych rezultatów pozwala na oszczędność czasu i środków w postępowaniu sądowym. Ważnym aspektem badań było dostosowanie metod walidacyjnych zaproponowanych modeli do sytuacji rzeczywistych spraw sądowych, tak by były one kompatybilne z praktyką i możliwe do bezpośredniej aplikacji.

Warto też wspomnieć, iż zaproponowane modele hybrydowe zostały przetestowane również do rozwiązania problemu porównawczego widm Ramana dla niebieskich lakierów samochodowych, co dodatkowo potwierdza ich uniwersalność. Natomiast użyteczność modeli LR była także testowana przez Autorkę w ramach interpretacji danych o mniejszej wymiarowości pochodzących z analiz niebieskich past długopisowych metodą mikrospektrofotometrii w zakresie widzialnym (MSP-Vis), próbek win włoskich opisanych przez 27 parametrów fizykochemicznych oraz stosunków izotopów ołowiu w próbkach szkieł wyznaczonych metodą spektrometrii mas stosunków izotopowych.