

UCZENIE MASZYNOWE – wstęp teoretyczny

Uczenie maszynowe (ML, ang. *Machine Learning*) jest gałęzią sztucznej inteligencji (AI, ang. *Artificial Intelligence*), której ideą jest umożliwienie maszynom wykonywania swoich zadań przy użyciu inteligentnych programów. Istnieje wiele definicji wyjaśniających czym jest *Machine Learning*. Jedna z nich mówi, że:

„Program komputerowy (maszyna) uczy się na podstawie doświadczenia E w odniesieniu do pewnej klasy zadań T i miary efektywności P , jeśli jego efektywność wykonywania zadania T (mierzona za pomocą P) poprawia się wraz z doświadczeniem T ”.

Wyróżnia się trzy główne kategorie algorytmów uczenia maszynowego: przez wzmacnianie (ang. *Reinforcement Learning/Deep Learning*), nienadzorowane (ang. *Unsupervised Learning*) oraz nadzorowane (ang. *Supervised Learning*).

Metoda **uczenia przez wzmocnienie** ma zastosowanie w sytuacjach, gdzie brak jest zarówno określonych parametrów wejściowych, jak i wyjściowych. Jej działanie opiera się na wykorzystywaniu obserwacji zbieranych na podstawie interakcji ze środowiskiem. Po każdej akcji, jedyną informacją zwrotną otrzymywaną przez maszynę jest tzw. sygnał wzmocnienia (pozytywny lub negatywny). Celem algorytmu jest podejmowanie działań w taki sposób, aby zmaksymalizować poprawność rezultatów. Do przykładów tej techniki należą Q-Learning oraz Monte Carlo.

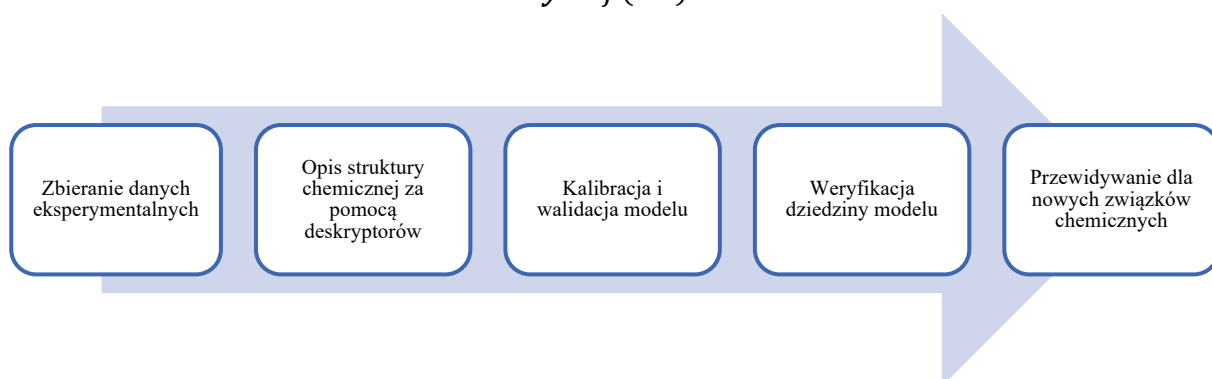
Uczenie nienadzorowane polega na dostarczeniu maszynie do nauki nieoznaczonego zbioru uczącego składającego się jedynie z cech/atrybutów wejściowych (wektor $x^{\vec{}}$) i niezawierającego etykiet. Algorytm samodoskonali się, wykorzystując duże, zróżnicowane zbiory danych. System uczy się modelu (funkcji h), którego zadaniem jest opisanie danych wyjściowych (np. wyszukanie prototypów, rozpoznawanie skupień). Przykładami tej kategorii są grupowanie metodą k-średnich i sieci neuronowe.

W **uczeniu nadzorowanym** algorytm tworzy funkcję przewidyującą przyszły wynik na podstawie danych wejściowych. Model predykcyjny otrzymuje instrukcję, czego ma się nauczyć i w jaki sposób. Dostarczany maszynie zestaw danych uczących zawiera zbiór cech/atrybutów wejściowych (wektor $x^{\vec{}}$) oraz poprawnie zaetykietowaną odpowiedź – wartość wyjściową ($f(x^{\vec{}})$). System uczy się modelu (funkcji h) mającego jak najlepiej aproksymować funkcję f w celu poprawnej predykcji etykiet przykładów, z którymi nie miał jeszcze styczności. Algorytmy uczenia nadzorowanego obejmują klasyfikację (przewidywanie dyskretnych odpowiedzi, np. SVM (ang. Support Machine Learning, k-najbliższych sąsiadów) oraz regresję (przewidywanie liczb rzeczywistych, np. regresja liniowa, drzewa decyzyjne). W uczeniu nienadzorowanym system szuka wzorców z bieżącego przykładu, natomiast w nadzorowanym korzysta z przykładów podanych wcześniej.

Przykładowym obszarem chemii komputerowej, w którym wykorzystywane są algorytmy uczenia nadzorowanego jest **QSAR/QSPR** (ang. *Quantitative Activity/Structure-Property Relationships*) – ilościowe modelowanie zależności pomiędzy strukturą chemiczną (ΔX) a aktywnością biologiczną/właściwością fizykochemiczną (Δy). Podstawą QSAR/QSPR jest założenie, że różnice w aktywności/właściwościach substancji chemicznych wynikają z różnic w ich budowie. W oparciu o zbiór obliczonych deskryptorów struktury, numerycznie wyrażających zmienność budowy chemicznej, oraz odpowiedni model matematyczny można interpolować brakujące informacje na temat związków, dzięki wartościom eksperymentalnym dostępnym dla podobnych do nich strukturalnie cząsteczek.

Ogólne równanie QSAR/QSPR wyrażone jest poniższym wzorem. Podstawowe kroki modelowania QSAR/QSPR przedstawiłam na **Rysunku 1**.

$$\Delta y = f(\Delta X)$$



Rysunek 1. Prosty schemat modelowania QSAR/QSPR.

Dopełnieniem modelowania QSAR/QSPR są różnorodne metody chemoinformatycznej, wielowymiarowej analizy danych (analiza i regresja głównych składowych, metoda częściowych najmniejszych kwadratów) oraz algorytmy sztucznej inteligencji (sieci neuronowe, algorytmy genetyczne). Połączenie wymienionych technik umożliwia modelowanie skomplikowanych zależności, tym samym stając się przyszłością badań naukowych nad powiązaniem aktywności/właściwości ze strukturą związków chemicznych oraz nadzieją na rozszerzenie obecnego stanu wiedzy.

METODA REGRESJI WIELOKROTNEJ (MLR)

Przygotowanie zbioru danych

Zgodnie z dobrą praktyką chemometryczną wyniki pomiarów do analizy problemów wieloparametrycznych (zależnych od dużej liczby zmiennych) powinny być gromadzone w odpowiednio zaplanowanych tabelach arkusza kalkulacyjnego. W przypadku modelowania QSAR należy skonstruować wektor \mathbf{y} (modelowana wielkość) oraz macierz \mathbf{X} (deskryptory). Dane zostały zorganizowane w oparciu o wzór macierzy przedstawiony na **Rysunku 2**.

Lek	Modelowana wielkość	Deskryptor 1	Deskryptor 2	Deskryptor 3	...	Deskryptor m
Lek i	y_i	X_{1i}	X_{2i}	X_{3i}	...	X_{mi}
Lek ii	y_{ii}	X_{1ii}	X_{2ii}	X_{3ii}	...	X_{mii}
Lek iii	y_{iii}	X_{1iii}	X_{2iii}	X_{3iii}	...	X_{miii}
...
Lek n	y_n	X_{1n}	X_{2n}	X_{3n}	...	X_{mn}

Rysunek 2. Wzór macierzy do modelowania QSAR.

Modelowanie QSAR wymaga podzielenia związków zbadanych eksperymentalnie na zbiór uczący (kalibracyjny), który wykorzystuje się do zbudowania modelu i jego walidacji wewnętrznej, oraz zbiór testowy (walidacyjny) służący do potwierdzenia zdolności prognostycznych modelu w procesie walidacji zewnętrznej. Najczęściej stosowanym algorytmem podziału jest 1:X polegający na posortowaniu wartości odpowiedzi (malejąco/rosnąco) i przypisaniu co x-tego obiektu do zbioru testowego, co zapewnia równomierny rozkład związków w zakresie modelowanej wielkości.

Autoskalowanie (standaryzacja) polega na transformacji zmiennych, aby wartość średnia każdej z nich była równa 0, a odchylenie standardowe równe jedności. W taki sposób wszystkie wymiary wielowymiarowej przestrzeni cech stają się współmierne. Autoskalowanie łączy w sobie centrowanie zmiennych względem początku układu współrzędnych oraz skalowanie przedziałowe, na skutek którego wyeliminowany zostaje wpływ jednostek użytych do pomiaru zmiennych. Wykonuje się je według wzoru:

$$z_{ij} = \frac{x_{ij} - x_j}{s_j}$$

gdzie: z_{ij} – standaryzowana wartość cechy,

x_{ij} – początkowa wartość zmiennej,

x_j – wartość średnia j-tej zmiennej,

s_j – odchylenie standardowe j-tej zmiennej.

Regresja wielokrotna (MLR, ang. *multiple linear regression*) to jedno z najprostszych, a zarazem najważniejszych narzędzi statystycznych, gdzie zmienna zależna (modelowana wielkość) wyrażana jest jako kombinacja liniowa zmiennych niezależnych (deskryptorów). W metodzie tej zakłada się model w postaci:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

gdzie: y – zmienna zależna (modelowana wielkość),

b_0 – wyraz wolny,

b_n – współczynniki modelu,

x_n – zmienne objaśniające (deskryptory).

Współczynniki dobierane są za pomocą metody najmniejszych kwadratów, czyli tak, aby suma różnic kwadratów pomiędzy wartościami przewidywanymi przez model a zmierzonymi eksperymentalnie (tzw. rezydualów) była jak najmniejsza. Wzór na współczynniki modelu regresyjnego przedstawia się następująco:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Po obliczeniu współczynników możliwe jest otrzymanie wektora \mathbf{y} zgodnie z równaniem:

$$\mathbf{y} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{H}\mathbf{y}$$

gdzie \mathbf{H} jest macierzą dźwigni wiążącą obliczone oraz eksperymentalne odpowiedzi modelu. Współczynniki dźwigni h (ang. *leverages*) są przydatne do sprawdzenia strukturalnej dziedziny modelu.

Istotnym ograniczeniem MLR w modelowaniu QSAR jest to, że gdy pomiędzy zmiennymi występują silne korelacje nie jest możliwe poprawne odwrócenie macierzy $(\mathbf{X}^T \mathbf{X})$, a więc wzór nie może zostać użyty do obliczenia współczynników \mathbf{b} . Występowanie takich korelacji w przypadku wielowymiarowych metod QSAR jest bardzo powszechne (niektóre deskryptory WHIM wyjaśniają tę samą cechę według różnych schematów ważenia, a więc są prawie całkowicie skorelowane), dlatego konieczne jest skorzystanie z innej metody np. PCR (ang. *Principal Component Regression*) lub dokonanie eliminacji jednego z pary skorelowanych zmiennych – zwykle o wyższej korelacji z innymi deskryptorami.

Miarę jakości dopasowania modelu (ang. *goodness-of-fit*) stanowią współczynnik determinacji (R^2 , ang. *determination coefficient*) oraz średni kwadratowy błąd kalibracji ($RMSE$, $RMSE_C$, ang. *Root Mean Square error of calibration*) obliczane według wzorów:

$$R^2 = 1 - \frac{\sum (y_{pred} - y_{obs})^2}{\sum (y_{obs} - \bar{y}_{obs})^2}$$

$$RMSE_C = \sqrt{\frac{\sum (y_{pred} - y_{obs})^2}{n_C}}$$

gdzie:

y_{pred} – przewidywana wartość zmiennej zależnej,

y_{obs} – obserwowana (eksperymentalna) wartość zmiennej zależnej,

n_C – liczebność zbioru kalibracyjnego.

Model QSAR uznaje się za dobrze dopasowany, gdy $R^2 > 0,65$ ($R^2 > 0,9$ w przypadku modelowania QSPR). R^2 zwiększa się wraz z dodawaniem kolejnych deskryptorów niezależnie od tego, czy redukują one niewyjaśnioną wariancję zmiennej zależnej.

Oceny statystycznej istotności przyjętego modelu dokonuje się z wykorzystaniem testu F-Snedecora – istotny, gdy informacja wyjaśniana przez model jest większa niż zawarta w jego błędach (im większa wartość tej statystyki, tym większa istotność modelu).

Dziedzina modelu

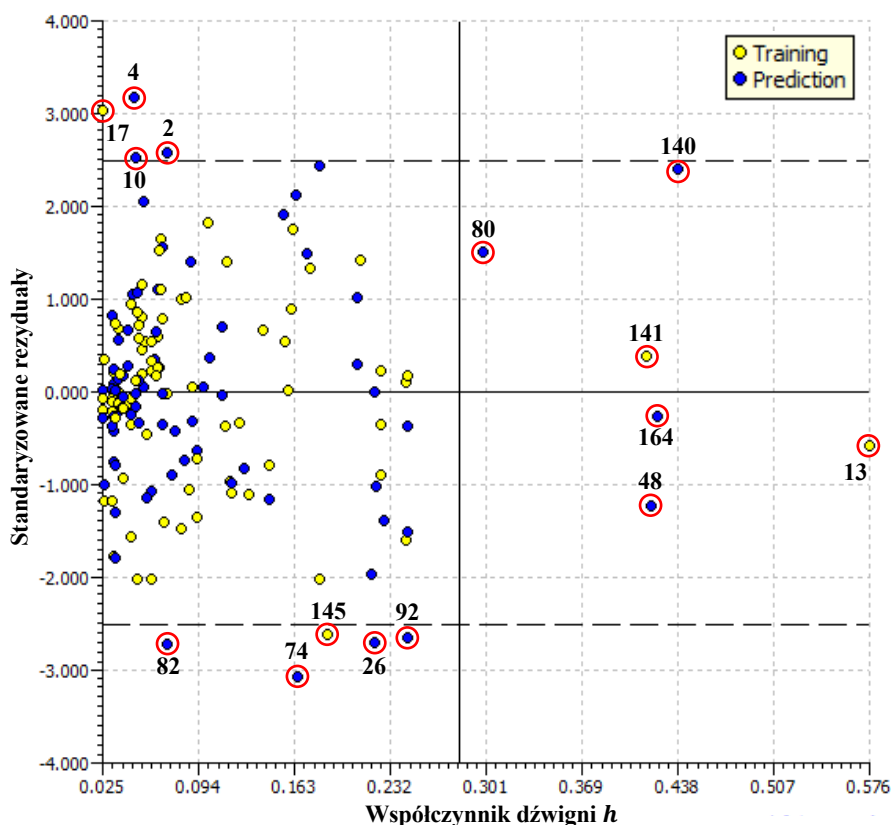
Dziedzina modelu (AD, ang. *Applicability Domain*) opisuje granice teoretycznego obszaru w przestrzeni cech związków, w którym przewidywania są wiarygodne. Przestrzeń tę wyznacza się poprzez zestawienie na osi odciętych współczynników dźwigni h , a na osi rzędnych wartości standaryzowanych rezydualów (wykres Williamsa). Współczynniki dźwigni określają strukturalne podobieństwo między związkiem a zbiorem uczącym. Wartość graniczna dźwigni jest wyrażana wzorem:

$$h^* = \frac{3p + 1}{n}$$

gdzie: p – liczba deskryptorów w modelu,

n – liczba związków w zbiorze uczącym.

Związki o wartościach $h > h^*$ rozpatrywane są jako leżące poza AD. Oś rzędnych przedstawia precyzję przewidywania przez model, w związku z czym wartości standaryzowanych rezydualów powinny znajdować się w przestrzeni ograniczonej przez $\pm 2,5/3$ odchylenia standardowe. Zdefiniowanie dziedziny modelu pozwala na sporządzenie mapy powierzchni błędu modelu w zależności od użytych deskryptorów oraz ocenę jego jakości.



Rysunek 3. Wykres Williamsa (MODEL 3.). $h^* = 0,282$.

Walidacja modelu

Elastyczność modelu QSAR (ang. *robustness*) to jego odporność na usunięcie jednego lub większej liczby związków ze zbioru uczącego. Weryfikuje się ją za pomocą walidacji krzyżowej (CV, ang. *cross-validation*). Zdolności prognostyczne modelu (ang. *predictive capability*) ocenia się przeprowadzając walidację zewnętrzną (ang. *external validation*).

Walidację krzyżową można określić mianem „walidacji wewnętrznej” (ang. *internal validation*), ponieważ wykorzystuje zbiór kalibracyjny również jako walidacyjny (niejednocześnie). Najprostszym typem walidacji krzyżowej jest „wyrzucić jeden” (LOO, ang. *leave-one-out*), który polega na wyłączeniu pojedynczych obiektów ze zbioru uczącego – stanowią one wówczas jednoelementowe zbiory walidacyjne. Na podstawie modelu skalibrowanego dla pozostałych $n - 1$ elementów przewiduje się wartość modelowanej wielkości dla wyrzuconego związku. Procedurę powtarza się n razy (kolejno dla każdego elementu), czego wynikiem jest n -elementowy zbiór wartości wyjść z modelu. W efekcie otrzymuje się zestaw n różnic pomiędzy prognoząmi modelu i odpowiedziami obiektów, które wykorzystywane są do wyznaczania współczynnika walidacji krzyżowej (Q_{cv}^2 , ang. *cross-validation determination coefficient*) oraz średniego kwadratowego błędu walidacji krzyżowej ($RMSE_{cv}$, ang. *root mean square error of cross-validation*) zgodnie z równaniami:

$$Q_{cv}^2 = 1 - \frac{\sum (y_{cv} - y_{obs})^2}{\sum (y_{obs} - \bar{y}_{obs})^2}$$

$$RMSE_{CV} = \sqrt{\frac{\sum (y_{cv} - y_{obs})^2}{n_C}}$$

gdzie:

y_{cv} – wartość zmiennej zależnej otrzymana z zastosowaniem walidacji krzyżowej,
 y_{obs} – obserwowana (eksperymentalna) wartość zmiennej zależnej,
 n_C – liczebność zbioru kalibracyjnego.

Istnieją również inne odmiany walidacji krzyżowej takie jak „wyrzuć więcej” (LMO, ang. *leave-more-out*), gdzie kolejno wyłącza się po kilka obiektów ze zbioru uczącego, a także metoda randomizowanej walidacji krzyżowej, w której elementy wyłączanych jednorazowo bloków wybierane są losowo.

W przeciwieństwie do walidacji krzyżowej, walidacja zewnętrzna dostarcza informacji dotyczących jakości przewidywania modelu posługując się zbiorem testowym. Miarą zdolności prognostycznych modelu QSAR są wartości współczynnika walidacji zewnętrznej (Q_{Ex}^2 , *external validation coefficient*) oraz średniego kwadratowego błędu przewidywania ($RMSE_{Ex}$, ang. *root mean square error of prediction*).

$$Q_{Ex}^2 = 1 - \frac{\sum (y_{pred} - y_{obs})^2}{\sum (y_{pred} - \bar{y}_{pred})^2}$$

$$RMSE_{Ex} = \sqrt{\frac{\sum (y_{pred} - y_{obs})^2}{n_T}}$$

gdzie:

y_{pred} – przewidywana wartość zmiennej zależnej,
 y_{obs} – obserwowana wartość zmiennej zależnej,
 n_T – liczebność zbioru walidacyjnego.

Dobry model QSAR powinien charakteryzować się możliwie bliskimi jedności wartościami R^2 , Q_{CV}^2 , Q_{Ex}^2 oraz porównywalnymi i możliwie małymi wartościami średnich błędów kwadratowych. Występowanie znacznych różnic pomiędzy wartościami błędów $RMSE_C$, $RMSE_{CV}$ i $RMSE_{Ex}$ wskazuje na zbyt duże podobieństwo związków należących do zbioru kalibracyjnego (ang. *overfitting*), a tym samym na małą zdolność modelu do generalizowania informacji.

