Algorytm imputacji brakujących wartości

Etap filtracji

Przez filtrację rozumiane jest usunięcie wierszy macierzy danych ***X***, w których liczba brakujących wartości przekracza zadany próg. Aby uniknąć usunięcia peptydów różnicujących, które występują tylko w jednej z porównywanych grup, filtracja poprzedzona jest testem badającym hipotezę zerową H_0 o braku zależności liczby brakujących wartości od przynależności do grupy. Statystyka testowa ma postać:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{J} \frac{\left(N_{ij} - E_{ij}\right)^2}{E_{ij}} \quad ,$$

gdzie N_{ij} są zaobserwowanymi liczbami brakujących ($i = 1$) i niebrakujących ($i = 2$) wartości w j-tej grupie, a E_{ji} są ich wartościami oczekiwanymi:

$$E_{ij} = \frac{N_i N_j}{N} \quad .$$

Przy prawdziwości H_0 statystyka testowa ma rozkład χ^2 o J–1 stopniach swobody. Wiersze, dla których wykryta zostanie istotna statystycznie (przy zadanym progu istotności) zależność pomiędzy brakującymi wartościami a przynależnością do jednej z grup badanych nie podlegają automatycznej filtracji, a dalsze postępowanie z nimi zależy od wyboru użytkownika i charakteru prowadzonej analizy. Domyślnie, wiersze te są odpowiednio odznaczane, a brakujące wartości zastępowane są minimalną wartością całego zbioru danych (taka rekonstrukcja dotyczy jedynie grupy próbek, w której wykryto najwięcej brakujących wartości). Możliwe jest jednak również włączenie tego typu cech do zbioru potencjalnie różnicujących albo zupełne wykluczenie ich z dalszej analizy.

Etap imputacji

Po filtracji w macierzy danych nadal będą występować pojedyncze brakujące wartości. W ich przypadku możliwe jest oczywiście trywialne postępowanie, polegające na wstawieniu pewnej wartości stałej, ale zdecydowanie lepszym rozwiązaniem wydaje się użycie metody odtworzenia brakujących wartości na podstawie pozostałych.

W badaniach z zakresu biologii molekularnej dużą popularność zdobyła, koncepcyjnie prosta, nieparametryczna metoda najbliższych sąsiadów (*K*NN – *K Nearest Neighbours*). Pierwotnie przeznaczona była ona dla danych pochodzących z mikromacierzy DNA, ale znalazła zastosowanie również w innych technikach pomiarowych. U jej podstaw leży obserwacja o istnieniu grup genów, których profile ekspresji wykazują znaczne podobieństwo. Chcąc oszacować

brakujący poziom ekspresji genu *i* w próbce *j*, szukamy *K* najbliższych mu genów (według pewnego ustalonego kryterium bliskości) spośród tych, dla których poziom ekspresji w próbce *j* został zmierzony prawidłowo. Następnie szukaną wartość $\tilde{x_{ij}}$ wyznaczamy jako średnią ważoną poziomów ekspresji w próbce *j* genów należących do wyznaczonego sąsiedztwa:

$$\tilde{x_{ij}} = \frac{\sum_{j=1}^{K} w_k x_{kj}}{\sum_{j=1}^{N} w_k} \quad .$$

W oryginalnej wersji metody wagi w_k były równe odwrotności odległości Euklidesa pomiędzy profilami ekspresji i stosowany był stały, wybrany z góry rozmiar sąsiedztwa.

Dla danych proteomicznych wymagane do działania algorytmu *K*NN założenie o występowaniu w zbiorze danych cech mających podobne profile ekspresji nie może budzić większych zastrzeżeń: nie dość, że istnieją grupy peptydów pochodzących z tych samych białek, to jeszcze część z nich jest reprezentowana przez więcej niż jeden stopień naładowania. Dlatego też w proponowanej metodzie na etapie imputacji brakujących wartości zastosowana została zmodyfikowana przez autora pracy wersja tego algorytmu.

Najistotniejsze z wprowadzonych modyfikacji dotyczą stosowanej miary odległości oraz sposobu w jaki budowane jest sąsiedztwo. Występująca w oryginalnej metodzie odległość Euklidesa zastąpiona została odległością wynikającą ze współczynnika korelacji liniowej pomiędzy wartościami ekspresji peptydów. Wagi w_k są więc wyznaczane jako:

$$w_k = \frac{1}{d_r} = \frac{1}{1 - r_{i,k}} \quad ,$$

gdzie $r_{i,k}$ jest współczynnikiem korelacji liniowej pomiędzy pomiędzy *i*-tym i *k*-tym wierszem macierzy danych ***X***. Ponieważ optymalna wielkość sąsiedztwa uwzględnianego podczas imputacji jest zależna od liczby cech i charakteru samych danych, lepszym rozwiązaniem wydaje się ustalanie jej w sposób dynamiczny i traktowanie liczby *K* jako maksymalnego dopuszczalnego rozmiaru otoczenia, do którego mogą jednak wejść jedynie cechy, których korelacja z aktualnie rekonstruowaną cechą jest większa od zadanego progu. W skrajnym przypadku, gdy liczba spełniających ten warunek cech jest równa 0, działanie algorytmu ograniczane jest do zastąpienia brakujących wartości średnią arytmetyczną pozostałych elementów tego samego wiersza macierzy ***X***. W odróżnieniu od swojego pierwowzoru, algorytm działa w sposób iteracyjny, korzystając z wyznaczonych w poprzedniej iteracji wartości na etapie budowania sąsiedztwa i określania wag wchodzących w jego skład cech. Proces zatrzymywany jest gdy średni kwadrat różnic pomiędzy imputowanymi wartościami z następujących po sobie iteracji spadnie poniżej nadanego progu.