Metody Bioinformatyki

Wykorzystanie PCA do analizy danych z mikromacierzy DNA

Maciej Czerniak, Marcin Kamionowski, Kacper Szkudlarek

23 stycznia 2012

Streszczenie

Dokumentacja końcowa realizacji projektu z przedmiotu "Metody Bioinformatyki". W ramach projektu wykonane zostało oprogramowanie wykorzystujące metodę Analizy Składowych Głównych (PCA) do przetwarzania danych uzyskanych z mikromacierzy DNA oraz wygenerowanych danych testowych.

1 Wstęp

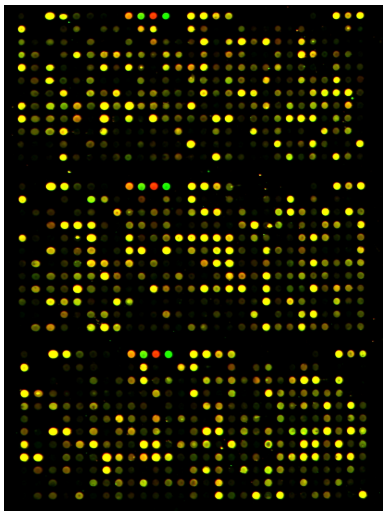
Analiza głównych składowych (ang. Principal Component Analysis, PCA) jest jedną ze statystycznych metod analizy czynnikowej. Zbiór danych składający się z N obserwacji, z których każda obejmuje K zmiennych, można interpretować jako chmurę N punktów w przestrzeni K-wymiarowej. Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd.. Tak przekształcone wartości współrzędnych nazywane są ładunkami wygenerowanych czynników (składowych głównych). W ten sposób konstruowana jest nowa przestrzeń obserwacji, w której najwięcej zmienności wyjaśniają początkowe czynniki. PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników.

Mikromacierz DNA jest to płytka szklana lub plastikowa z naniesionymi w regularnych pozycjach mikroskopowej wielkości polami (ang. spots), zawierającymi różniące się od siebie sekwencją fragmenty DNA. Fragmenty te są sondami, które wykrywają przez hybrydyzację komplementarne do siebie cząsteczki DNA lub RNA.

Dane (Rys: 1) uzyskiwane w eksperymentach prowadzonych z wykorzystaniem mikromacierzy to wartości intensywności czerwonej oraz zielonej fluorescencji każdego z pół na płytce. Jedonorazowo w eksperymencie możliwe jest badanie ekspresji kilku tysięcy genów, dlatego uzyskane dane są wysoce złożone i wielowymiarowe.

2 Wykonana implementacja

Implementacja zgodnie z opisem zawartym w dokumentacji wstępnej została podzielona na kilka części:



Rysunek 1: Pokolorowana próbka danych z mikromacierzy cDNA

1. Modułu wczytywania danych.

   Dane do analizy dostarczone są do aplikacji w surowej postaci plików tekstowych (tabbed separate data). Przyjęty format analizowanych danych zakłada, że w kolumnach zostały oznaczone poszczególne przeprowadzone eksperymenty, natomiast w wierszach znajdują się cechy, które poddane zostaną redukcji przy użyciu algorytmu PCA. Przy wczytywaniu użytkownik może wybrac jaki zakres danych ma być poddany analizie. Możliwe jest pominięcie pewnej ilości początkowych cech, a także wybranie zakresu eksperymentów, które maja być brane pod uwagę. Dane zostają wczytane do struktury danych Mat z biblioteki OpenCV [1], która następnie przekazywana jest do analizy.

2. Stworzenie graficznego interfejsu użytkownika.

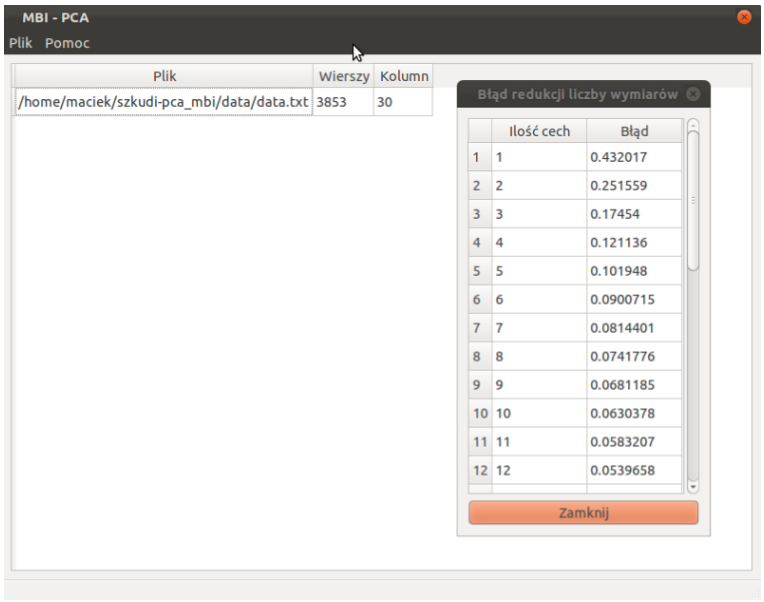
   Interfejs graficzny(Rys: 2) został stworzony w oparciu o wieloplatformowa bibliotekę Qt [2]. Z poziomu interfejsu możliwe jest wczytywanie danych, ustawienie parametrów importu wczytywanego pliku, wyświetlenie błędu średniokwadratowego.

3. Moduł analizujący i przetwarzający dane wejściowe.

   Analiza danych polega na poddaniu ich analizie składowych głównych (PCA). Wykorzystany został moduł PCA znajdujący się w bibliotece OpenCV. Dane poddawane są analizie w postaci struktury danych Mat i w takiej samej formie są zwaracane. Dane poddawane są także odwrotnemu PCA w celu odtworzenia oryginalnej macierzy. Dzięki temu możliwe staje się policzenie błedu średniokwadratowego, dla różnej liczby redukowanych składowych.

4. Moduł generatora danych

   Dodatkowo w celu wygenerowania danych testowych stowrzony został moduł generatora. Generowne dane są opisem jednostajengo prostolinowego ruchu obiektu w przestrzeni, widzianego z wielu punktów obserwacji (kamer). W celu wygenerowania



Rysunek 2: Graficzny interfejs użykownika programu

danych należy podać pozycję początkową i końcowa ruchu, ilość i pozycję kamer, a także ilość przeprowadonych (co 1s) pomiarów. Wygeneroweane dane zapisywane są do pliku tekstowego w postaci kompatybilnej z modułem wczytywania danych.

Całość implementacji została wykonana w języku C/C++ pod kontrolą systemu Linux. Dzięki użyciu wieloplatformowych bibliotek (Qt, OpenCV) aplikacja jest w pełni przenośna. Uruchomienie programu na nowym systemie wymaga jedynie rekompilacji pod kontrolą tego systemu. Do synchronizacji pracy programistów w czasie tworzenia projektu, użyty został rozproszony system wersjonowania plików Git.

Literatura

[1] Strona domowa projektu opencv: http://opencv.willowgarage.com/wiki/.

[2] Strona domowa projektu qt: http://qt.nokia.com/.