Projekt zaliczeniowy

Biologia obliczeniowa

Analiza danych mikromacierzowych

* **Cele**

Celem zadania było stworzenie klasyfikatora binarnego stwierdzającego stan zdrowia względem danej choroby dla danych mikromacierzowych zawierających próbki od osób zdrowych i chorych oraz znalezienie zależności między wynikami a dodatkowymi informacjami o pacjentach. Dodatkowym celem było znalezienie genów, które są najbardziej istotne przy klasyfikacji.

* **Dane**

Zbiór danych mikromacierzowych pochodzących od 107 osób, w tym 49 zdrowych i 58 chorych na raka płuc. Dane zawierają dodatkowe informacje na temat stadium choroby, płci, nałogów.

* **Środowisko i zasoby**

Obliczenia zostały wykonane w środowisku R, wykorzystując następujące biblioteki:

* Biobase
* GEOquery
* kernlab
* kohonen

Do wczytania i przetwarzania danych wykorzystaliśmy kroki kursu „Microarray analysis course” <http://www.bigre.ulb.ac.be/Users/jvanheld/statistics_bioinformatics/practicals/microarrays_berry_2010/>

* **Podejście**

W przyjętym przez nas podejściu zostały zrealizowane następujące kroki:

* Odczytanie danych z pliku .soft pobranego z bazy GEO
* Normalizacja danych
* Znormalizowaliśmy dane do przedziału [0,1] przez podzielenie każdej wartości dla danego genu przez maksymalną wartość. W ten sposób tracimy informację o natężeniu genów, ale nie powinna być ona potrzebna.
* Problem wyboru cech

Ponieważ używany przez nas zbiór danych był bardzo duży (ponad 22 tys. genów), należało dokonać wyboru genów (cech), które będą potrzebne do klasyfikacji i pozbyć się danych nieistotnych.

Ze względu na tak dużą liczbę danych, niektóre ze sposobów „feature selection” były niemożliwe do zrealizowania na posiadanym sprzęcie i środowisku (złożoność pamięciowa była zbyt duża).

Wybór polegał na sprawdzeniu, które geny nie różnią się między grupą zdrowych a chorych osób. Porównywaliśmy wariancję/średnią dla wartości w obu grupach i porównywaliśmy je ze sobą. Geny, dla których różnica nie przekraczała ustalonego progu, zostały odrzucone. Do tego celu zastosowaliśmy następujące algorytmy:

* Test F-Snedecoera – sprawdzenie jednorodności wariancji dwóch prób.
* Test Wilcoxona dla różnych wariancji - sprawdzanie równości średnich w przypadku pozytywanego wyniku testu F-S.
* Test T-Studenta – porównanie średnich dla równych wariancji z testu F-S.
* Wybrane zostały geny otrzymane z dwóch ostatnich testów.
* Klasyfikacja

Do podziału danych na dwie grupy wykonaliśmy obliczenia na danych przed i po filtrowaniu KPCA.

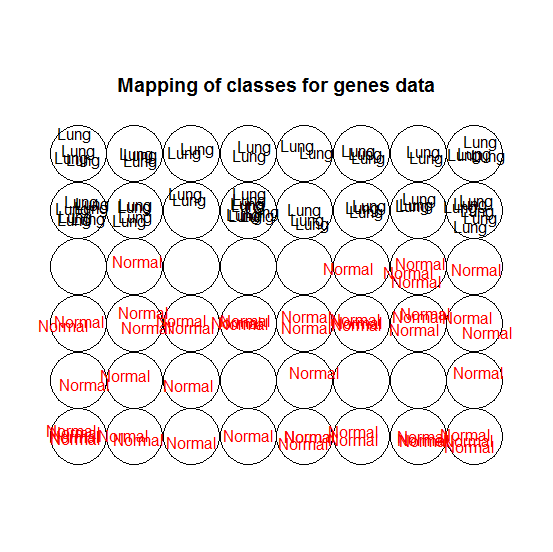
* XYF – sieć samoorganizująca się z nadzorowanym nauczaniem do klasyfikacji i wizualizacji podziału
* SVM (Support Vector Machine) – do podziału danych wielowymiarowych na dwie klasy
* Klastrowanie i wybór najbardziej znaczących genów

W celu znalezienia najbardziej istotnych grup genów, dokonaliśmy klastrowania (podziału na grupy ze względu na podobieństwo) za pomocą mapy samoorganizującej. Następnie wybraliśmy grupę, dla której zróżnicowanie wśród osób zdrowych i chorych było największe.

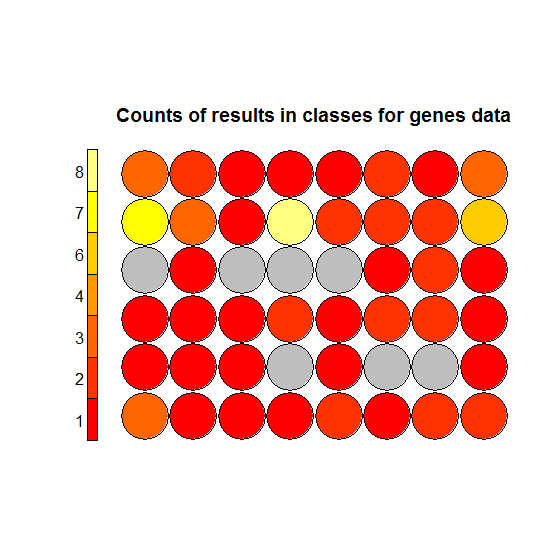
* **Wyniki**

Wyniki w postaci wykresów i tabela genów są automatycznie zapisane do pliku w lokalizacji roboczej.

* Klasyfikacja SOM



|  |  |  |
| --- | --- | --- |
| Wykres nr | Opis | Dane wejściowe do sieci |
| 1 | Wykres przedstawia wynik klasyfikacji wyników badań pacjentów za pomocą nauczania nadzorowanego sieci SOM. Czarne wyniki odpowiadają tkankom chorym, czerwone – zdrowym. | Każdy wektor danych wejściowych odpowiada wynikom dla jednej próbki od danego pacjenta, kolejne elementy wektora to wartości genów wybranych w poprzedniej fazie wyboru cech. |

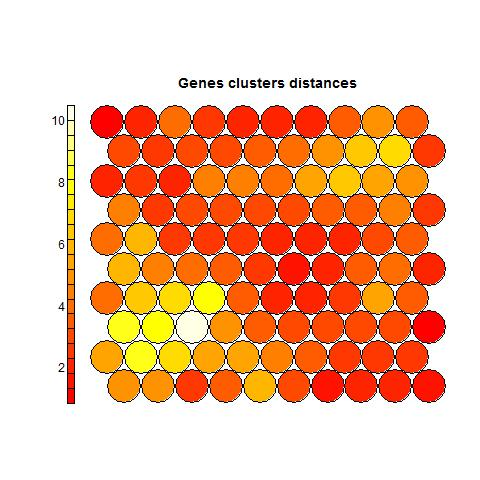


|  |  |  |
| --- | --- | --- |
| Wykres nr | Opis | Dane wejściowe do sieci |
| 2 | Wykres przedstawia liczbę wyników zaklasyfikowanych do każdej klasy w sieci SOM. | Każdy wektor danych wejściowych odpowiada wynikom dla jednej próbki od danego pacjenta, kolejne elementy wektora to wartości genów wybranych w poprzedniej fazie wyboru cech. |

*Dla nauczania z nadzorowaniem w używanym pakiecie „Kohonen” nie jest dostępny wykres odległości klas (tylko dla nauczania bez nadzoru)*

* Klasyfikacja SVM

Dla wybranych genów udało się wytrenować SVM z zerowym błędem nauczania.

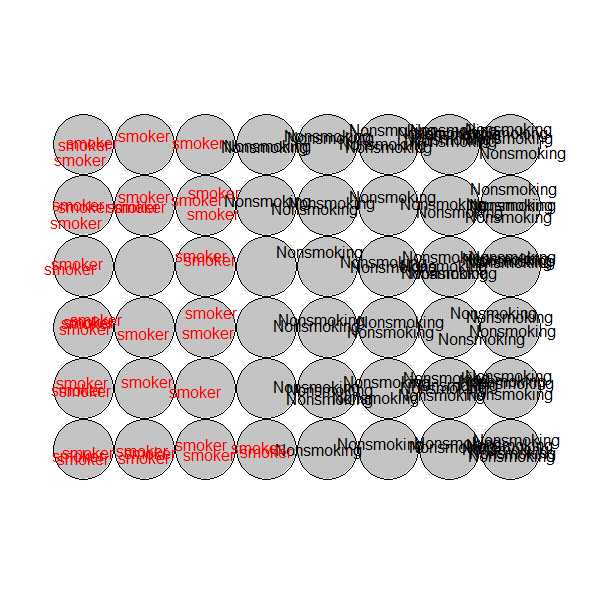
* Klastrowanie

|  |  |  |
| --- | --- | --- |
| Wykres nr | Opis | Dane wejściowe do sieci |
| 3 | Wykres przedstawia odległości między poszczególnymi klastrami genów otrzymanych za pomocą sieci SOM | Każdy wektor danych wejściowych odpowiada wartościom jednego z genów wyselekcjonowanych w poprzedniej fazie dla każdego z pacjentów |

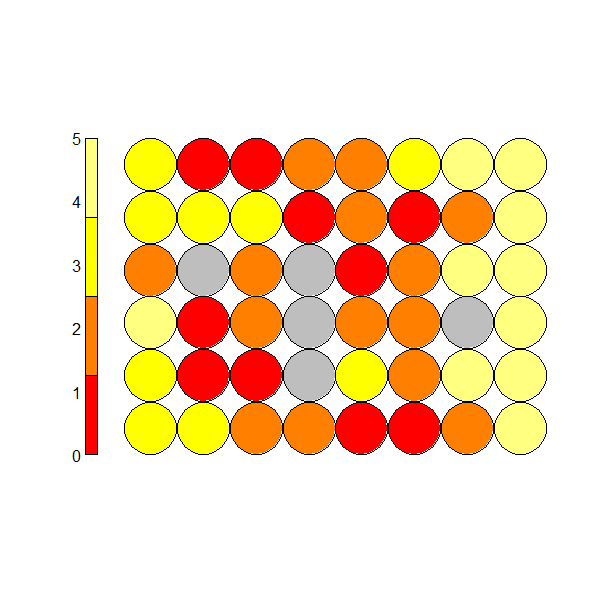
Geny należące do „zwycięskiego klastra” znajdują się w dołączonym pliku *most\_significant\_genes.txt.*

* Klasyfikacja SOM dla palaczy/niepalących

Przyjąwszy analogiczną metodologię, jak w przypadku podziału na chorych i zdrowych dowiedliśmy istnienia korelacji między paleniem a zmianami genowymi prowadzącymi do raka płuc. Odzwierciedlają to poniższe wykresy.



|  |  |  |
| --- | --- | --- |
| Wykres nr | Opis | Dane wejściowe do sieci |
| 4 | Wykres przedstawia klasyfikację wyników badań genetycznych dla osób palących i niepalących. Dane oznaczone jako „smoker” pochodzą od palaczy, „Nonsmoker” – od osób niepalących. | Każdy wektor danych wejściowych odpowiada wynikom dla jednej próbki od danego pacjenta, kolejne elementy wektora to wartości genów wybranych w poprzedniej fazie wyboru cech. |



|  |  |  |
| --- | --- | --- |
| Wykres nr | Opis | Dane wejściowe do sieci |
| 5 | Wykres przedstawia liczbę wyników zaklasyfikowanych do każdej klasy w sieci SOM. | Każdy wektor danych wejściowych odpowiada wynikom dla jednej próbki od danego pacjenta, kolejne elementy wektora to wartości genów wybranych w poprzedniej fazie wyboru cech. |

* **Ewaluacja wyników**

W celu ewaluacji otrzymanych rezultatów, wykorzystaliśmy dane medyczne (informacje na temat genów, które są oznaczane w testach na prawdopodobieństwo wystąpienia raka). Jako ewaluację metody traktowaliśmy liczbę pozostałych genów oraz liczbę genów, które powinny zostać i są obecne w zbiorze. Te geny to:

* CHRNA3 (210221\_AT)
* CHRNA5 (206533\_AT)
* CHRNB4 ( 207516\_AT)
* GSTP1 (200824\_AT)
* GSTM1 ( 204550\_X\_AT)
* ALK (208211\_S\_AT,208212\_S\_AT)
* EML4 (220386\_S\_AT)

Niestety w grupie genów wybranych w procesie selekcji nie pojawiają się pożądane dane. Być może wybrana metoda wyboru cech nie była wystarczająco dokładna.

Klasyfikacja danych pacjentów okazała się skuteczna.