Wybrane metody statystyczne w selekcji genów

Idzi Siatkowski, Alicja Szabelska, Joanna Zyprych

Katedra Metod Matematycznych i Statystycznych, UP w Poznaniu

9 grudnia 2009

Plan prezentacji

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp
- Wybrane testy statystyczne w środowisku R

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp
- Wybrane testy statystyczne w środowisku R
- Przedstawienie danych

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp
- Wybrane testy statystyczne w środowisku R
- Przedstawienie danych
- Przedstawienie metod i narzędzi

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp
- Wybrane testy statystyczne w środowisku R
- Przedstawienie danych
- Przedstawienie metod i narzędzi
- Wyniki

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp
- Wybrane testy statystyczne w środowisku R
- Przedstawienie danych
- Przedstawienie metod i narzędzi
- Wyniki
- Wnioski

Plan prezentacji

- Cel pracy, metody i narzędzia - wstęp
- Wybrane testy statystyczne w środowisku R
- Przedstawienie danych
- Przedstawienie metod i narzędzi
- Wyniki
- Wnioski
- Literatura

Cel pracy, metody i narzędzia - wstęp

Cel pracy, metody i narzędzia - wstęp

- przedstawienie testów używanych do selekcji genów

Cel pracy, metody i narzędzia - wstęp

- przedstawienie testów używanych do selekcji genów
- sprawdzenie zależności: test - metoda predykcji

Cel pracy, metody i narzędzia - wstęp

- przedstawienie testów używanych do selekcji genów
- sprawdzenie zależności: test - metoda predykcji
- wnioskowanie o użyteczności danego testu

Cel pracy, metody i narzędzia - wstęp

- przedstawienie testów używanych do selekcji genów
- sprawdzenie zależności: test - metoda predykcji
- wnioskowanie o użyteczności danego testu
- 3 zestawy danych: leukemia72, ovarian, ostra białaczka szpikowa

Cel pracy, metody i narzędzia - wstęp

- przedstawienie testów używanych do selekcji genów
- sprawdzenie zależności: test - metoda predykcji
- wnioskowanie o użyteczności danego testu
- 3 zestawy danych: leukemia72, ovarian, ostra białaczka szpikowa
- zaimplementowanie w R: 5 testów statystycznych i 3 metody predykcji

Cel pracy, metody i narzędzia - wstęp

- przedstawienie testów używanych do selekcji genów
- sprawdzenie zależności: test - metoda predykcji
- wnioskowanie o użyteczności danego testu
- 3 zestawy danych: leukemia72, ovarian, ostra białaczka szpikowa
- zaimplementowanie w R: 5 testów statystycznych i 3 metody predykcji
- porównanie na diagramach Venna wyników kolejnych testów

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )
Fligner-Killeen test	stats	fligner.test( )

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )
Fligner-Killeen test	stats	fligner.test( )
Bartlett test	stats	bartlett.test( )

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )
Fligner-Killeen test	stats	fligner.test( )
Bartlett test	stats	bartlett.test( )
Brown-Forsyth test	HH	hov( )

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )
Fligner-Killeen test	stats	fligner.test( )
Bartlett test	stats	bartlett.test( )
Brown-Forsyth test	HH	hov( )

Kontrola FDR

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )
Fligner-Killeen test	stats	fligner.test( )
Bartlett test	stats	bartlett.test( )
Brown-Forsyth test	HH	hov( )

Kontrola FDR

- p.adjust(...., method="BH") - korekta Benjaminiego-Hochberga (1995)

Wybrane testy statystyczne w środowisku R

test	pakiet	funkcja
F test	DEDS	deds.stat.linkC( )
	stats	oneway.test(...,var.equal=TRUE)
	genefilter	rowFtest( )
	golub	mt.teststat(..., test="f")
Kruskal-Wallis test	stats	kruskal.test( )
Fligner-Killeen test	stats	fligner.test( )
Bartlett test	stats	bartlett.test( )
Brown-Forsyth test	HH	hov( )

Kontrola FDR

- p.adjust(...., method="BH") - korekta Benjaminiego-Hochberga (1995)
- mt.rawp2adjp(multtest)

Dane Leukemia72

T.R.Golub (1999)

Dane Leukemia72

T.R.Golub (1999)

- *http://www.genome.wi.mit.edu/MPR*

Dane Leukemia72

T.R.Golub (1999)

- *http://www.genome.wi.mit.edu/MPR*
- ALLBcell & ALLTcell & AML (przewlekłe i ostra białaczka szpikowa)

Dane Leukemia72

T.R.Golub (1999)

- *http://www.genome.wi.mit.edu/MPR*
- ALLBcell & ALLTcell & AML (przewlekłe i ostra białaczka szpikowa)
- macierze oligonukleotydowe Affymetrix

Dane Leukemia72

T.R.Golub (1999)

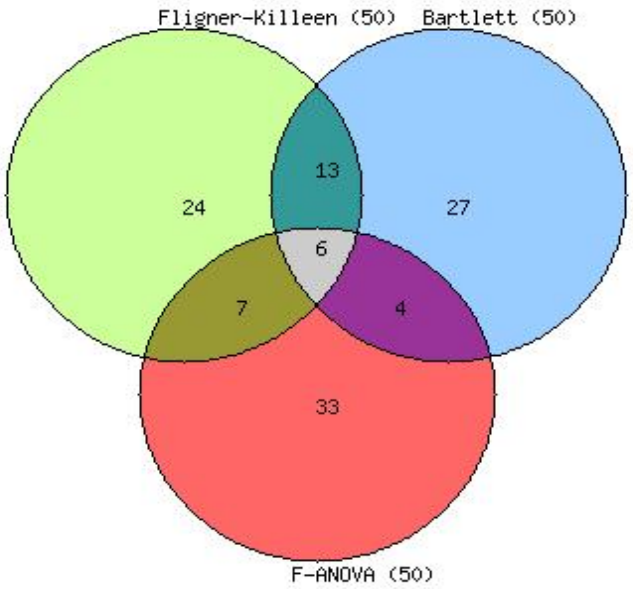
- *http://www.genome.wi.mit.edu/MPR*
- ALLBcell & ALLTcell & AML (przewlekłe i ostra białaczka szpikowa)
- macierze oligonukleotydowe Affymetrix
- 7129 genów

Dane Leukemia72

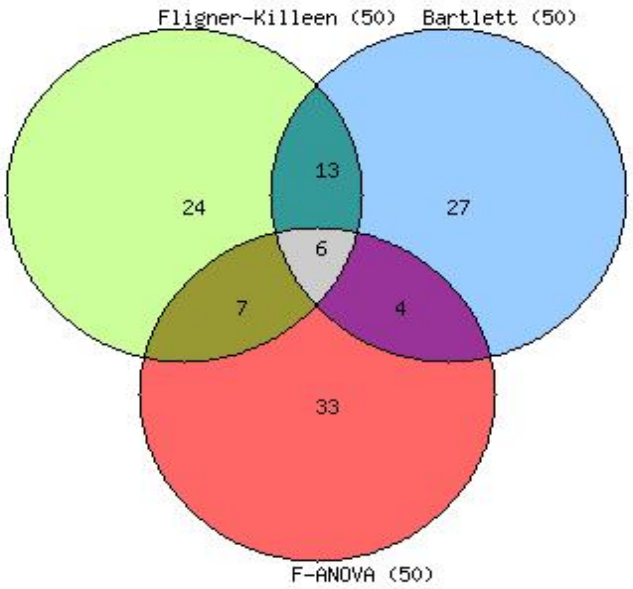
T.R.Golub (1999)

- *http://www.genome.wi.mit.edu/MPR*
- ALLBcell & ALLTcell & AML (przewlekłe i ostra białaczka szpikowa)
- macierze oligonukleotydowe Affymetrix
- 7129 genów
- 72 próby: 38 ALLBcell, 9 ALLTcell, 25 AML

Porównanie testów statystycznych

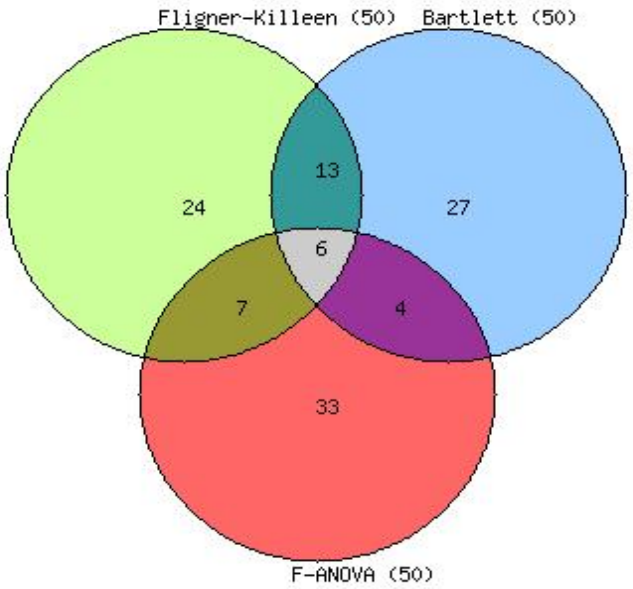


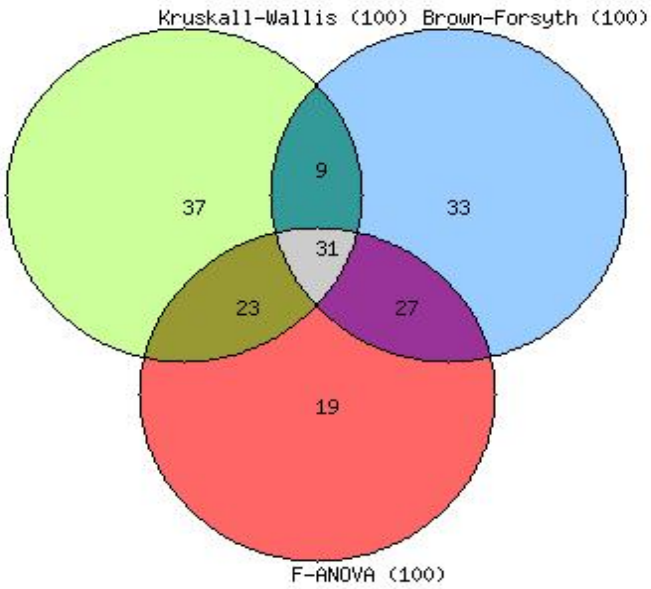
Porównanie testów statystycznych



50 genów informatywnych

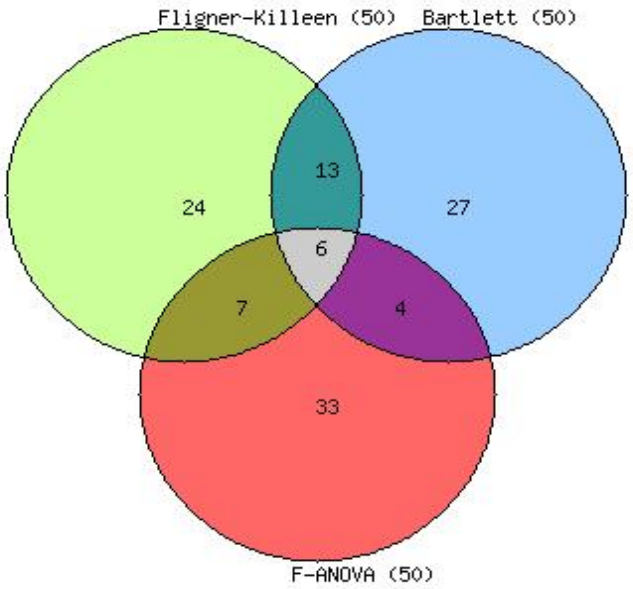
Porównanie testów statystycznych



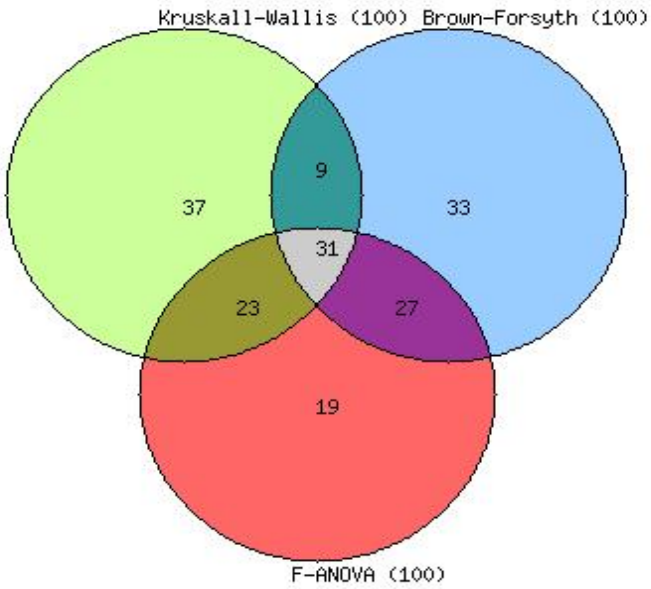


50 genów informatywnych

Porównanie testów statystycznych



50 genów informatywnych



100 genów informatywnych

Klasyfikatory

Metody predykcji

Klasyfikatory

- naiwny Bayesowski (NB)

Metody predykcji

Klasyfikatory

- naiwny Bayesowski (NB)
- najbliższych sąsiadów (KNN)

Metody predykcji

Klasyfikatory

- naiwny Bayesowski (NB)
- najbliższych sąsiadów (KNN)
- wektorów nośnych (SVM)

Metody predykcji

Klasyfikatory

- naiwny Bayesowski (NB)
- najbliższych sąsiadów (KNN)
- wektorów nośnych (SVM)

Zaimplementowanie klasyfikatorów w R

Metody predykcji

Klasyfikatory

- naiwny Bayesowski (NB)
- najbliższych sąsiadów (KNN)
- wektorów nośnych (SVM)

Zaimplementowanie klasyfikatorów w R

- pakiet **MLInterfaces**

Metody predykcji

Klasyfikatory

- naiwny Bayesowski (NB)
- najbliższych sąsiadów (KNN)
- wektorów nośnych (SVM)

Zaimplementowanie klasyfikatorów w R

- pakiet **MLInterfaces**
- `MLearn(L ~ ., data = dane, method=c(NaiveBayesI,knnI,svmI),training_set)`

Walidacja wyników

LOOCV walidacja krzyżowa

1. liczba prób n
2. próba ucząca n – 1
3. klasyfikator = NaiveBayesI,kNN,SVM(dane-1)
4. sprawdzenie klasyfikacji dla n-tej próby
5. błąd - liczba złych klasyfikacji w n powtórzeniach

Funkcja w R: **LOOCV(bestglm) - autorzy: A.I. McLeod i C. Xu**

Rezultaty-błąd predykcji

test	klasyfikator	50 genów	100 genów	200 genów
Kruskall-Wallis	NB	3	3	1
	KNN	4	7	3
	SVM	2	4	3
	średnia	3	4.7	2.3

Rezultaty-błąd predykcji

test	klasyfikator	50 genów	100 genów	200 genów
Kruskall-Wallis	NB	3	3	1
	KNN	4	7	3
	SVM	2	4	3
	średnia	3	4.7	2.3
Brown-Forsyth	NB	0	0	1
	KNN	3	4	3
	SVM	1	2	5
	średnia	1.3	2	3

Rezultaty-błąd predykcji

test	klasyfikator	50 genów	100 genów	200 genów
Kruskall-Wallis	NB	3	3	1
	KNN	4	7	3
	SVM	2	4	3
	średnia	3	4.7	2.3
Brown-Forsyth	NB	0	0	1
	KNN	3	4	3
	SVM	1	2	5
	średnia	1.3	2	3
F-ANOVA	NB	1	1	1
	KNN	3	5	4
	SVM	3	3	3
	średnia	2.3	3	2.7

Rezultaty - c.d.

test	klasyfikator	50 genów	100 genów	200 genów
Flingern-Killeen	NB	2	3	1
	KNN	6	5	3
	SVM	7	4	4
	średnia	5	4	2.7

Rezultaty - c.d.

test	klasyfikator	50 genów	100 genów	200 genów
Flingern-Killeen	NB	2	3	1
	KNN	6	5	3
	SVM	7	4	4
	średnia	5	4	2.7
Bartlett	NB	3	3	0
	KNN	6	5	3
	SVM	9	10	13
	średnia	6	6	5.3

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA
- max błąd predykcji – Fligner–Killeen, Bartlett

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA
- max błąd predykcji – Fligner–Killeen, Bartlett
- kNNI – zwiększona ilość błędów predykcji

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA
- max błąd predykcji – Fligner–Killeen, Bartlett
- kNNI – zwiększona ilość błędów predykcji
- NaiveBayesI – zmnieszona ilość błędów predykcji

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA
- max błąd predykcji – Fligner–Killeen, Bartlett
- kNNI – zwiększona ilość błędów predykcji
- NaiveBayesI – zmnieszona ilość błędów predykcji
- SVM – zła predykcja dla testu Bartletta

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA
- max błąd predykcji – Fligner–Killeen, Bartlett
- kNNI – zwiększona ilość błędów predykcji
- NaiveBayesI – zmnieszona ilość błędów predykcji
- SVM – zła predykcja dla testu Bartletta
- największa liczba informatywnych genów wspólnych: Brown-Forsyth, F–ANOVA

Wnioski

- min błąd predykcji – Brown-Forsyth, F-ANOVA
- max błąd predykcji – Fligner–Killeen, Bartlett
- kNNI – zwiększona ilość błędów predykcji
- NaiveBayesI – zmnieszona ilość błędów predykcji
- SVM – zła predykcja dla testu Bartletta
- największa liczba informatywnych genów wspólnych: Brown-Forsyth, F–ANOVA
- test Bartletta – najmniejsza liczba informatywnych genów wspólnych przy porównaniach Brown–Forsyth, F–ANOVA, Kruskall –Wallis

Literatura

- Dechang Chen, Zhenqiu Liu, Xiaobin Ma, Dong Hua. Selecting Genes by Test Statistics. Journal of Biomedicine and Biotechnology. 2005; 2: 132–138.
- Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors Rusing gene expression data. J Amer Statist Assoc. 2002; 97(457): 77–87.
- Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439): 531–537.
- Hartung J, Argac D, Makambi KH. Small sample properties of tests on homogeneity in oneway ANOVA and meta-analysis. Statist Papers. 2002; 43: 197–235.
- Krzyśko M., Wołyński W., Górecki T., Skorzybut M. Systemy uczące się. Rozpoznawanie wzorców analiza skupień i redukcja wymiarowości. Wydawnictwo Naukowo Techniczne. 2008.
- Welsh JB, Zarrinkar PP, Sapinoso LM, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci USA. 2001; 98(3): 1176–1181.

Dziękuję za uwagę