Warianty modelowania (supervised, semisupervised, soft i unsupervised learning) z użyciem mieszanin rozkładów normalnych oraz przykłady zastosowania w analizach danych bioinformatycznych

Przemyslaw.Biecek@gmail.com

Toruń 4 IV 2011

Plan referatu

1. Wprowadzenie do modeli mieszanin Gaussowskich
2. Pomiędzy supervised a unsupervised
3. Algorytm EM
4. Problem wyboru modelu
5. Pakiet dla programu R: bgmm
6. Przykładowe zastosowania w bioinformatyce
7. Interesujące kierunki rozwoju

Wprowadzenie do modeli mieszanin Gaussowskich

Rozważmy parę zmiennych losowych (X, Y), gdzie realizacje zmiennej X to wektory z przestrzeni $\mathcal{R}^d$ a realizacje zmiennej Y to liczby całkowite ze zbioru $\{1, ..., k\}$. Przyjmijmy tymczasowo, że k jest znane.

O zmiennej Y zakładamy, że ma rozkład wielomianowy z parametrami $(\pi_1, ..., \pi_k)$, a rozkład warunkowy zmiennej X to wielowymiarowy rozkład normalny $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma_y)$.

Rozkład łączny przedstawić można jako

$$f(x, y) = \pi_y f(x, \theta_y),$$

gdzie $\theta_y = (\mu_y, \Sigma_y)$ a $f(x, \theta_y)$ to gęstość rozkładu normalnego o parametrach θ_y w punkcie x.

Zmienna Y określa indeks komponenty mieszaniny rozkładów normalnych a X to zmienna wylosowana z danego komponentu.

Pomiędzy supervised a unsupervised

Interesujący nas mechanizm losowy generuje realizacje zmiennej losowej (X, Y). Rozważać możemy różne warianty w zależności od tego co badacz obserwuje.

- Uczenie z nadzorem (Klasyfikacja/*Fully supervised modeling*), badacz obserwuje realizacje zarówno zmiennej X jak i Y,
- Uczenie bez nadzoru (Analiza skupisk/*Unsupervised modeling*), badacz obserwuje wyłącznie realizacje zmiennej X, zmienna Y jest nieobserwowana,
- Uczenie z częściowym nadzorem (*Semi-supervised modeling*), badacz obserwuje realizacje zmiennej X i część realizacji zmiennej Y (to które realizacje obserwujemy jest niezależne od X),
- Uczenie z rozmytym nadzorem (*belief-based / soft-label modeling*), badacz obserwuje realizacje zmiennej X i pewną, być może niepoprawną informację o część realizacji zmiennej Y.

Uczenie z nadzorem

Obserwujemy n par (x_i, y_i) gdzie $i \in \{1, ..., n\}$.

Znamy k ponieważ jest to liczba różnych wartości przyjmowanych przez zmienną Y.

Jeżeli interesuje nas klasyfikacja to zagadnienie to sprowadza się do liniowej lub kwadratowej klasyfikacji, w zależności od tego czy założymy, że $\forall_i \Sigma_i = \Sigma$.

Uczenie bez nadzoru

Obserwujemy m wektorów (x_i) gdzie $i \in \{1, ..., m\}$.

Możemy rozważać sytuacje w których znamy lub w których nie znamy k.

Jeżeli interesuje nas analiza skupisk to problem sprowadza się do *model based clustering*.

Uczenie z częściowym nadzorem

Obserwujemy n par (x_i, y_i) gdzie $i \in \{1, ..., n\}$ i m wektorów x_i gdzie $i \in \{m+1, ..., m+n\}$.

Możemy rozważać sytuacje w których znamy lub w których nie znamy k (na podstawie obserwowanych Y wiemy jakie jest minimalne k).

W zastosowaniach spotykać można przypadki w których proporcje m do n są bardzo różne.

Uczenie z rozmytym nadzorem

Obserwujemy n par (x_i, b_i) gdzie $i \in \{1, ..., n\}$ i m wektorów x_i gdzie $i \in \{m+1, ..., m+n\}$.

Wektory $b_i = (b_{i,1}, ..., b_{i,k}) \in \mathcal{R}^k$, dodatkowo aby ułatwić interpretację zakładamy, że $\sum_j b_{i,j} = 1$. Wektory b_i są znane i opisują rozkład prawdopodobieństwa opisujący nasze przekonania (*belief-based*) lub wagi (*soft-label*) przynależności do klas.

W algorytmie *soft-label* przyjmuje się, że jeżeli klasa nie jest znana to wszystkim obserwacjom przypisuje się równe wagi, w algorytmie *belief-based* nie przypisuje się żadnych wag.

Różnica polega na interpretacji wag, przełoży się też na wyniki.

Pomiędzy supervised a unsupervised

Zagadnienia, które będą nas interesowały:

- Estymacja parametrów rozkładu wektora π_k i θ_k,
- Wybór d o ile nie jest znane,
- Przyporządkowanie nowych lub istniejących obserwacji do najbardziej prawdopodobnych komponent.

Algorytm EM

Praca uważana za pierwszą kompletne sformułowanie algorytmu EM.
A. P. Dempster, N. M. Laird, and D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, 39:1-38, 1977.

Popularny algorytm do wyznaczana estymatorów ML gdy (z różnych powodów) w modelu występują zmienne nieobserwowane.

Iteracyjnie powtarzamy kroki:

- Krok E, wyznaczamy

$$Q(\theta|\theta^{(n)}) = E[\log p(Z|\theta)|y, \theta^{(n)}]$$

- Krok M, maksymalizujemy $Q(\theta|\theta^{(n)})$ po θ.

Algorytm EM dla uczenia bez nadzoru

Funkcja wiarogodności

$$l(\mathcal{X}, B, \Phi) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j f(x_i, \theta_j) \right).$$

Krok E w iteracji q

$$t_{i,j}^{(q+1)} = \pi_j f\left(x_i, \theta_j^{(q)}\right) / \sum_{l=1}^{k} \pi_l f\left(x_i, \theta_l^{(q)}\right).$$

Krok M w iteracji q

$$\pi_j^{(q+1)} = \sum_{i=m+1}^{n} t_{i,j}^{(q+1)} / (n-m),$$

$$\mu_j^{(q+1)} = \left(\sum_{i=1}^{n} x_i t_{i,j}^{(q+1)} \right) / \left(\sum_{i=1}^{n} t_{i,j}^{(q+1)} \right),$$

$$\Sigma_j^{(q+1)} = \left(\sum_{i=1}^{n} \left(x_i - \mu_j^{(q+1)}\right)^T \left(x_i - \mu_j^{(q+1)}\right) t_{i,j}^{(q+1)} \right) / \left(\sum_{i=1}^{n} t_{i,j}^{(q+1)} \right).$$

Algorytm EM dla „belief-based”

Funkcja wiarogodności

$$l(\mathcal{X}, B, \Phi) = \sum_{i=1}^{m} \log \left(\sum_{j=1}^{k} b_{i,j} f(x_i, \theta_j) \right) + \sum_{i=m+1}^{n+m} \log \left(\sum_{j=1}^{k} \pi_j f(x_i, \theta_j) \right),$$

Krok E w iteracji q

$$t_{i,j}^{(q+1)} = \begin{cases} b_{i,j} f\left(x_i, \theta_j^{(q)}\right) / \sum_{l=1}^{k} b_{i,l} f\left(x_i, \theta_l^{(q)}\right) & i \leqslant M \\ \pi_j f\left(x_i, \theta_j^{(q)}\right) / \sum_{l=1}^{k} \pi_l f\left(x_i, \theta_l^{(q)}\right) & i > M \end{cases}$$

Krok M w iteracji q

$$\pi_j^{(q+1)} = \sum_{i=m+1}^{n} t_{i,j}^{(q+1)} / (n-m),$$

$$\mu_j^{(q+1)} = \left(\sum_{i=1}^{n} x_i t_{i,j}^{(q+1)} \right) / \left(\sum_{i=1}^{n} t_{i,j}^{(q+1)} \right),$$

$$\Sigma_j^{(q+1)} = \left(\sum_{i=1}^{n} \left(x_i - \mu_j^{(q+1)} \right)^T \left(x_i - \mu_j^{(q+1)} \right) t_{i,j}^{(q+1)} \right) / \left(\sum_{i=1}^{n} t_{i,j}^{(q+1)} \right).$$

Algorytm EM dla „soft-label”

Funkcja wiarogodności

$$l(\mathcal{X}, P, \Phi) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p_{i,j} \pi_k f(x_i, \theta_j) \right).$$

Krok E w iteracji q

$$t_{i,j}^{(q+1)} = p_{i,j} \pi_j^{(q)} f\left(x_i, \theta_j^{(q)}\right) / \sum_{l=1}^{k} p_{i,l} \pi_l^{(q)} f\left(x_i, \theta_l^{(q)}\right),$$

Krok M w iteracji q

$$\pi_j^{(q+1)} = \sum_{i=1}^{n} t_{i,j}^{(q+1)} / n,$$

$$\mu_j^{(q+1)} = \left(\sum_{i=1}^{n} x_i t_{i,j}^{(q+1)} \right) / \left(\sum_{i=1}^{n} t_{i,j}^{(q+1)} \right),$$

$$\Sigma_j^{2\ (q+1)} = \left(\sum_{i=1}^{n} \left(x_i - \mu_j^{(q+1)} \right)^T \left(x_i - \mu_j^{(q+1)} \right) t_{i,j}^{(q+1)} \right) / \left(\sum_{i=1}^{n} t_{i,j}^{(q+1)} \right).$$

Inicjacja algorytmu

Jak się okazuje inicjacja parametrów dla algorytmu EM ma istotny wpływ na zbieżność

- inicjacja tylko na podstawie obserwacji z etykietą,
- inicjacja na podstawie wszystkich obserwacji używając algorytmów uczenia bez nadzoru, następnie uzgadnianie etykiet z znalezionymi komponentami.

Uzgadnianie etykiet można wykonać heurystyką (może prowadzić do minimów lokalnych) lub rozpatrując wszystkie permutacje etykiet (niewykonalne dla dużych zbiorów etykiet).

Problem wyboru modelu

Wykorzystujemy kryterium GIC (*Generalized Information Criteria*)

$$GIC(M) = -2L(M|X) + p|M|,$$

gdzie p kara za wielkość modelu ($p = 2$ AIC, $p = log(n)$ BIC).

- Wybór liczby komponentów d,
- Wybór struktury μ, Σ (*mean, within, between, cov*).

struktura	# parametrów	struktura	# parametrów
DDDD	$kd + kd(d+1)/2$	EDDD	$d + kd(d+1)/2$
DDD0	$kd + kd$	EDD0	$d + kd$
DDED	$kd + 2k$	EDED	$d + 2k$
DDE0	$kd + k$	EDE0	$d + k$
DEDD	$kd + d(d+1)/2$	EEDD	$d + d(d+1)/2$
DED0	$kd + d$	EED0	$d + d$
DEED	$kd + 2$	EEED	$d + 2$
DEE0	$kd + 1$	EEE0	$d + 1$

Oprogramowanie

Istniejące oprogramowanie w popularnie używanych pakietach statystycznych

- Uczenie z nadzorem (klasyfikacja), lda() i qda() w pakiecie MASS,
- Uczenie bez nadzoru (analiza skupisk), Mclust() w pakiecie mclust,
- Uczenie z częściowym nadzorem (semisupervised), pakiet phyclust dla R, Spider dla Matlaba.
  Dla problemu *semisupervised* dużo narzędzi dla metody SVM, mniej dla *Gaussian mixtures*.

Pakiet dla programu R: bgmm

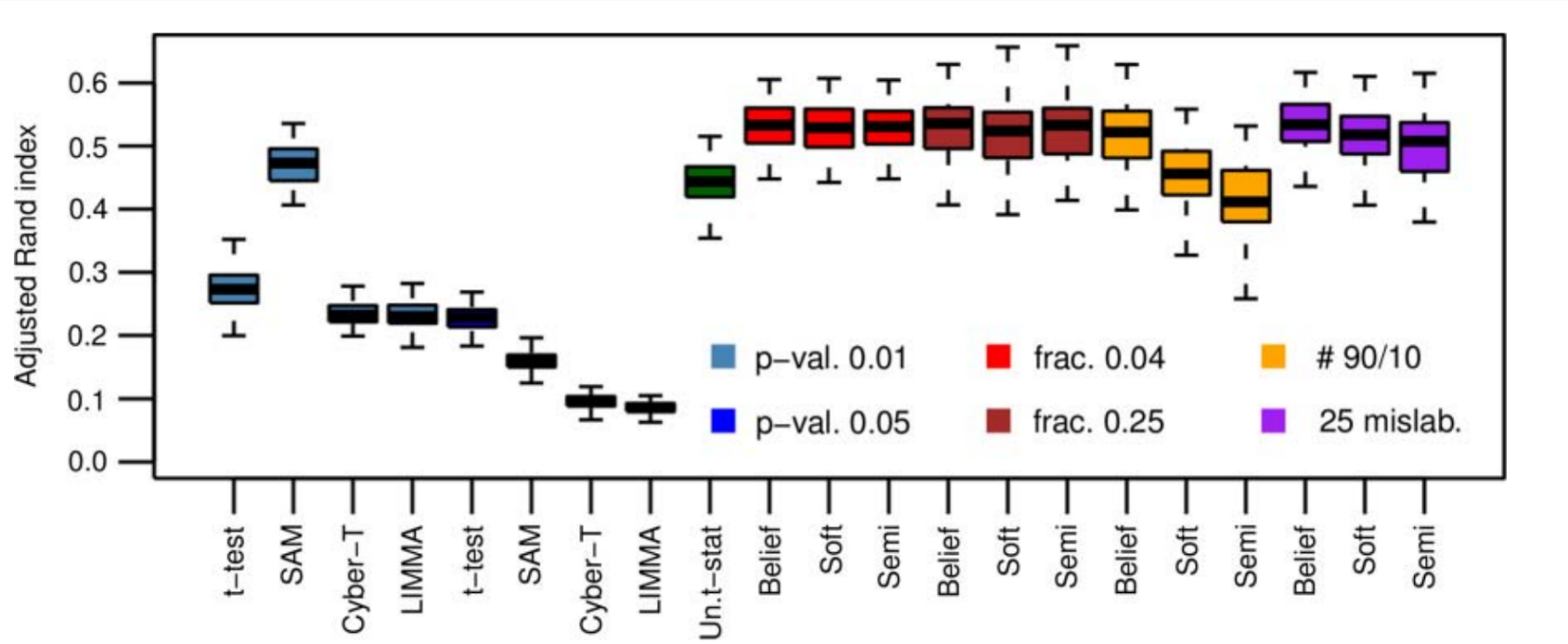
Opracowaliśmy pakiet bgmm dla programu R.
Skrót bgmm pochodzi od belief-based Gaussian mixture modeling.

Zalety/cechy/funkcjonalność

- wybór (ręczny lub automatyczny) liczby komponentów Gaussowskich,
- wybór (ręczny lub automatyczny) struktury średnich i macierzy wariancji dla komponentów Gaussowskich,
- implementacja pięciu wersji modeli mieszanin Gaussowskich (supervised, semisupervised, belief, soft, unsupervised),
- narzędzia do wizualizacji modeli i losowania danych.

Przykładowe zastosowania w bioinformatyce

Na danych symulacyjnych sprawdzaliśmy jak radzą sobie różne metody w zależności od różnych parametrów.



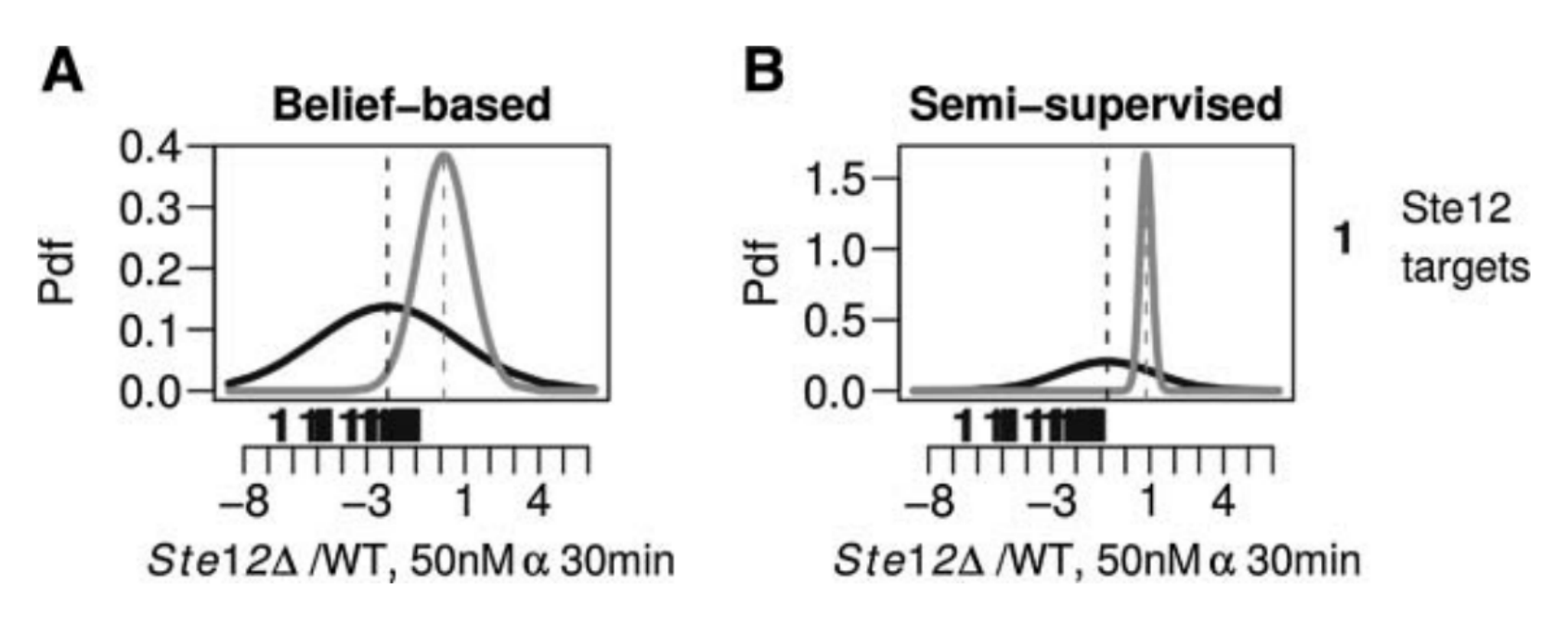
Przykładowe zastosowania w bioinformatyce

Interesuje nas znalezienie genów których odpowiedź na podanie feromonu jest regulowana przez czynnik transkrypcyjny Ste12.

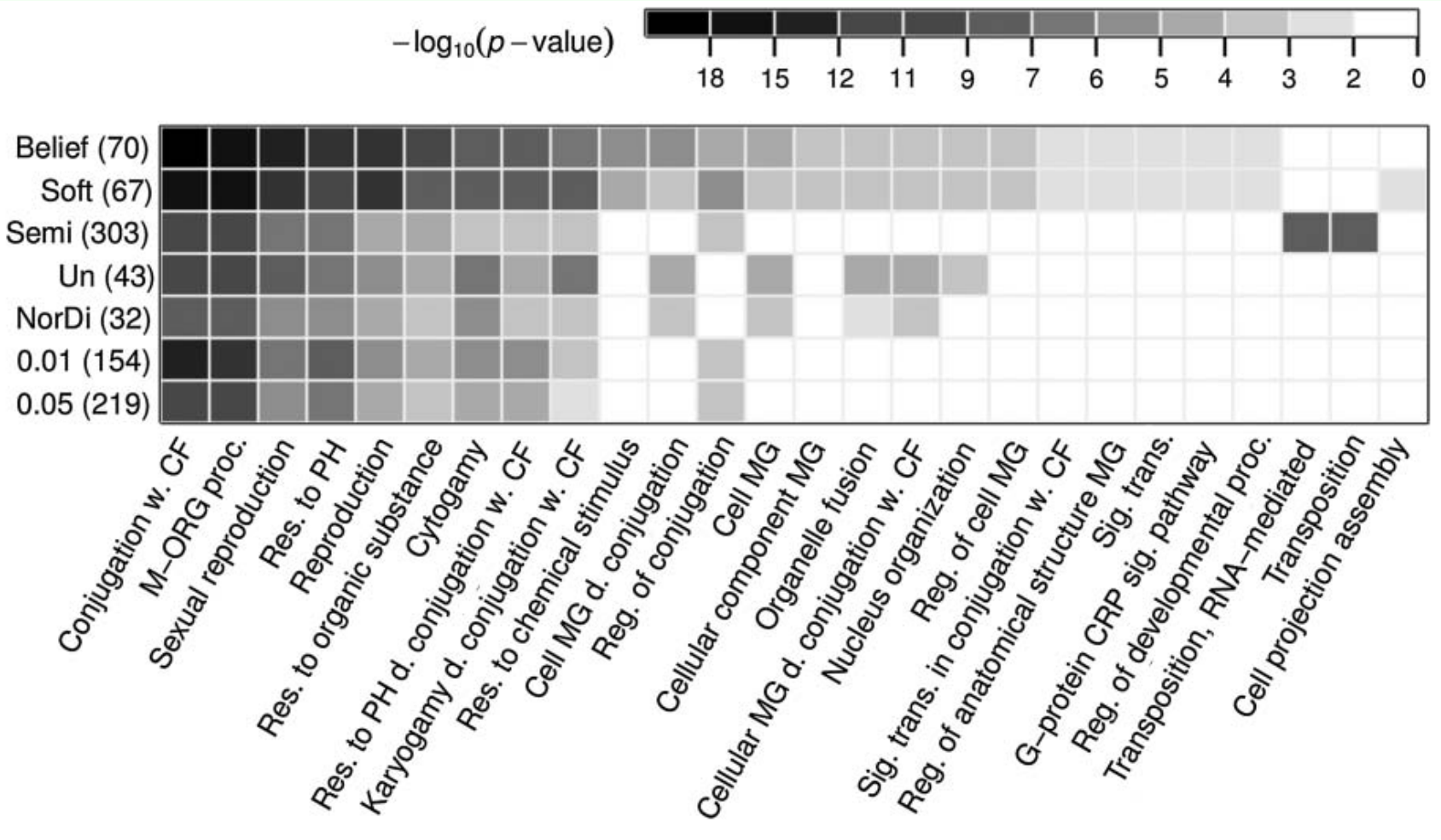
Dla wybranych 602 genów drożdży sprawdzamy jak zmienia się ekspresja genów w komórkach potraktowanych 50nM roztworem feromonu w zależności od tego czy komórki drożdży mają *knock-out* genu kodującego czynnik transkrypcyjny Ste12.

Spodziewamy się, że odpowiedź na feromon będzie inna dla genów w których usunięto Ste12 i dla *wild-type*. Z innych badań mamy informację o genach do których promotorów wiąże się czynnik transkrypcyjny Ste12. Są to dla nas przykłady danych z etykietami, czyli genów co do których jesteśmy przekonani, że powinny inaczej zareagować.

Przykładowe zastosowania w bioinformatyce



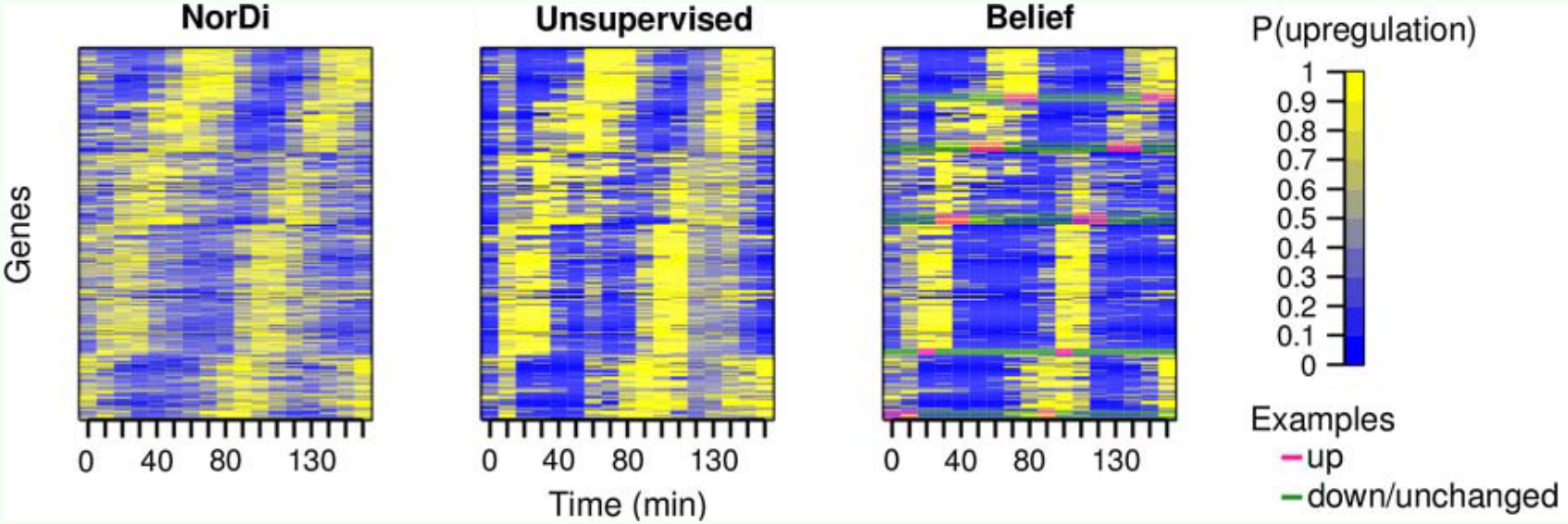
Przykładowe zastosowania w bioinformatyce



CF, cellular fusion; M-ORG, multi-organism; Res., response; PH, pheromone; MG, morphogenesis; Reg., regulation; CRP, coupled receptor protein; Sig. trans., signal transduction; w., with; d., during.

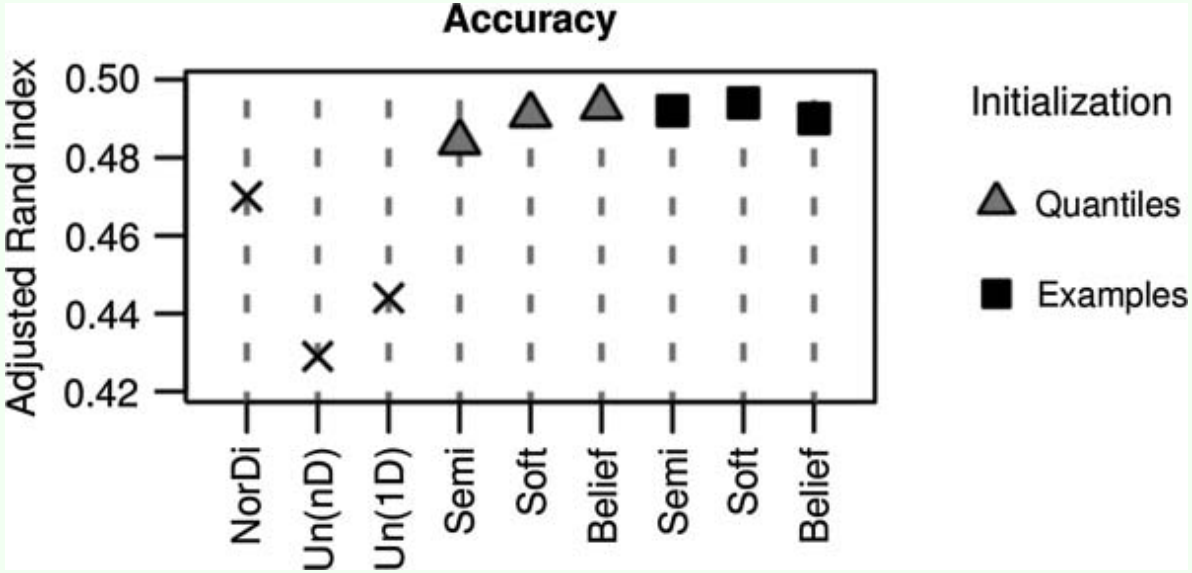
Przykładowe zastosowania w bioinformatyce

Analiza genów biorących udział w cyklu komórkowym. Wyróżnionych jest pięć faz cyklu, informację o etykietach można uzyskać na podstawie danych GO.



Przykładowe zastosowania w bioinformatyce

Adjusted Rand index określa zgodność dwóch przyporządkowań do skupisk.



Podsumowanie i interesujące kierunki rozwoju

- Możliwość uwzględnienia informacji (niepewnej) dotyczącej części obserwacji,
- Efektywna implementacja, odpowiednia dla problemów często spotykanych w bioinformatyce.

- Bardzo dużo danych, za dużo by zmieścić w pamięci, za dużo by zastosować jakikolwiek iteracyjny algorytm na całych danych.
  Kierunek: Algorytmy klasy stochastyczny równoległy gradient, „*online learning*" dla EM.
- Komponenty mają strukturę, nie są płaskie.
  Kierunek: Modyfikacja funkcji wiarogodności i probabilistyczny opis ontologii dla komponentów.

References

1. Szczurek E, Biecek P, Tiuryn J, Vingron M (2010). „Introducing knowledge into differential expression analysis". J Comput Biol, 17(8), 953-67.
2. Biecek P, Szczurek E, Tiuryn J, Vingron M (2011). „The R package bgmm: mixture modeling with uncertain knowledge". Journal of Statistical Software (in prep).
3. Come E, Oukhellou L, Denux T, Aknin P (2009). „Learning from partially supervised data using mixture models and belief functions". Pattern Recognition, 42(3), 334-348.
4. Fraley C, Raftery AE (2006). „MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering". Technical Report 504, University of Washington.
5. Steele R, Raftery AE (2009). „Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models". Technical Report 559, University of Washington.
6. Zhu X, Goldberg AB (2009). „Introduction to Semi-Supervised Learning. Morgan Claypool Publishers".