

Echo + баесовская кластеризация

Дмитрий Грошев, Ирина Горбунова



СПбАУ

04.06.2012

Краткое содержание предыдущих серий

- ▶ <http://uc-echo.sourceforge.net/>
- ▶ эффективно исправляет ошибки

- ▶ улучшенная версия утилиты hammer
- ▶ кластеризует k-меры перед исправлением ошибок
- ▶ эффективнее начальной версии

- ▶ попытка улучшить ECHO
- ▶ кластеризация похожих ридов

Как работает ECHO

2 стадии

- ▶ поиск пересекающихся ридов (соседей)
- ▶ исправление ошибок



- ▶ поиск точно совпадающего k-мера hashmap'ом

$$K = \max\left(\frac{L}{5}, \log(\sqrt[4]{N * L})\right)$$

- ▶ подбор ω и ϵ (покрытие конкретного нуклеотида должно иметь распределение Пуассона)
- ▶ поиск неточных соседей по ω и ϵ

- ▶ В группе соседей делается MAP по каждому нуклеотиду и confusion matrix
- ▶ На основе исправленных ридов пересчитывается confusion matrix
- ▶ Повторять до схождения

Это EM-алгоритм

Maximum A Posteriori

$$P(A_i) = \prod_{j=1}^N \Phi_{READS_{j,i}, A}$$

$$\arg \max_{X=A,T,G,C} P(X)$$

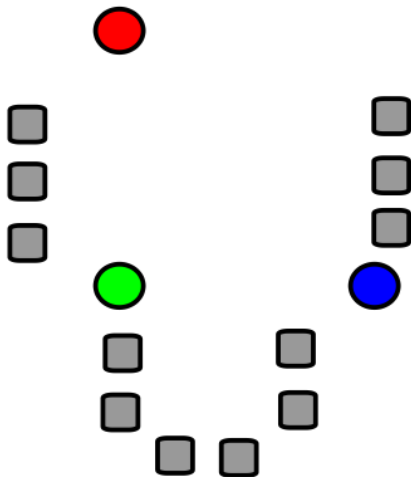
Что сделали мы



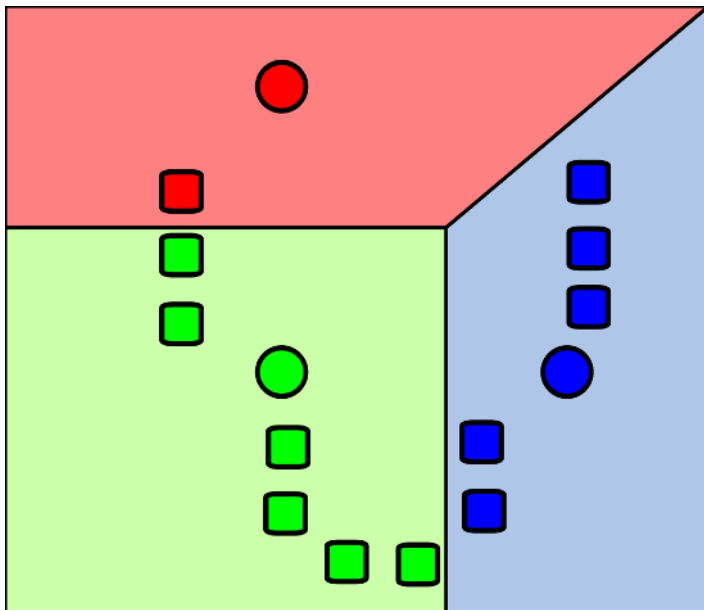
- ▶ выбираются произвольные центры
- ▶ для каждой точки ищется ближайший центр (диаграмма Вороного)
- ▶ пересчитываются центры

Это тоже EM

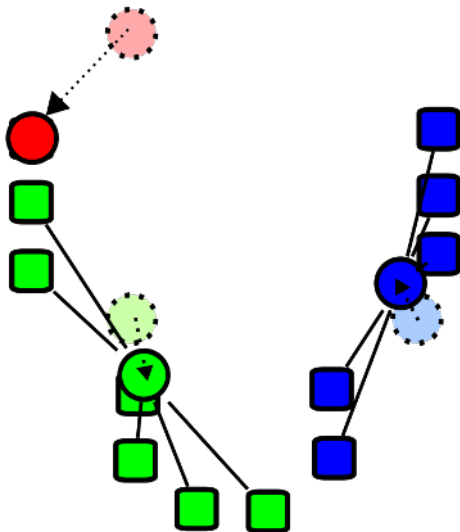
K-Means



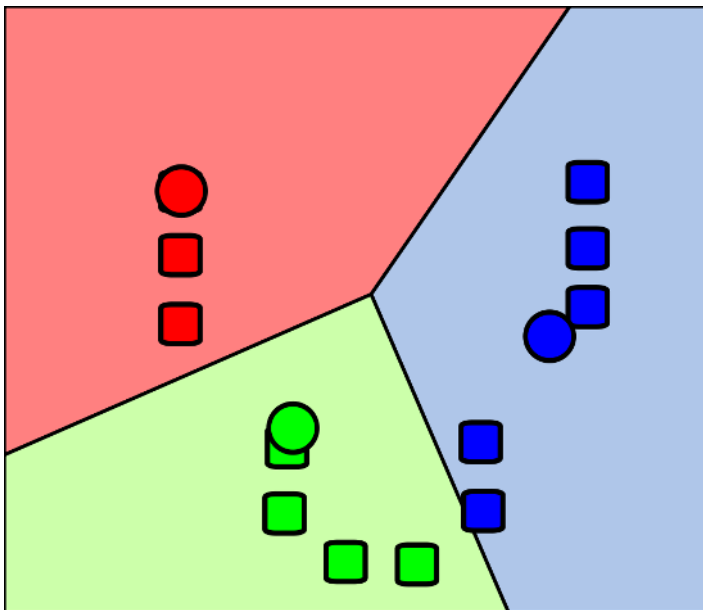
K-Means



K-Means



K-Means



Модификация K-Means с более сложной инициализацией

- ▶ центроидом выбирается случайная точка
- ▶ подсчитывается расстояние от каждой точки до ближайшего центроида
- ▶ новый центроид выбирается из оставшихся точек с вероятностью, пропорциональной квадрату расстояния от остальных центроидов

Риды не полностью совпадают (в отличие от k-меров):

```
ATGCATGCA
  TCATGTACGG
  TGCATGCAC
```

Добавим спецсимвол:

```
ATGCATGCA__  
__TCATGTACGG  
_TGCATGCAC__
```

Расстояние от `_` до любого символа = 0.

Если выбран неполный центроид, дополнить из ближайших (случайный нуклеотид пропорционально расстоянию)

```
___ATGC  
GCGCTGC  
GCGCTAC  
GCGCT___ <---  
TATATAT  
TCTAT___  
TATGTAT
```

```
___ATGC
GCGCTGC *
GCGCTAC *
GCGCT___ < - - -
TATATAT
TCTAT___
TATGTAT
```



```
  ___ATGC  
GCGCTGC  
GCGCTAC  
[GCGCTGC] <---  
TATATAT  
TCTAT___  
TATGTAT
```

Расстояние = $1 - \text{likelihood}$

Снова MAP:

group1	group2
___ATGC	TATATAT
GCGCTGC	TCTAT__
GCGCTAC	TATGTAT
GCGCT__	
-----	-----
GCGCTGC	TATATAT

- ▶ кластеризация кластеризует
- ▶ запускается
- ▶ на больших наборах данных ещё не запускали

Вопросы?