Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Phương pháp k láng giềng
K nearest neighbors

Đỗ Thanh Nghị
dtnghi@cit.ctu.edu.vn

Cần Thơ
02-12-2008

Nội dung

- Giới thiệu về KNN
- Kết luận và hướng phát triển

Nội dung

- **Giới thiệu về KNN**
- Kết luận và hướng phát triển

K nearest neighbors

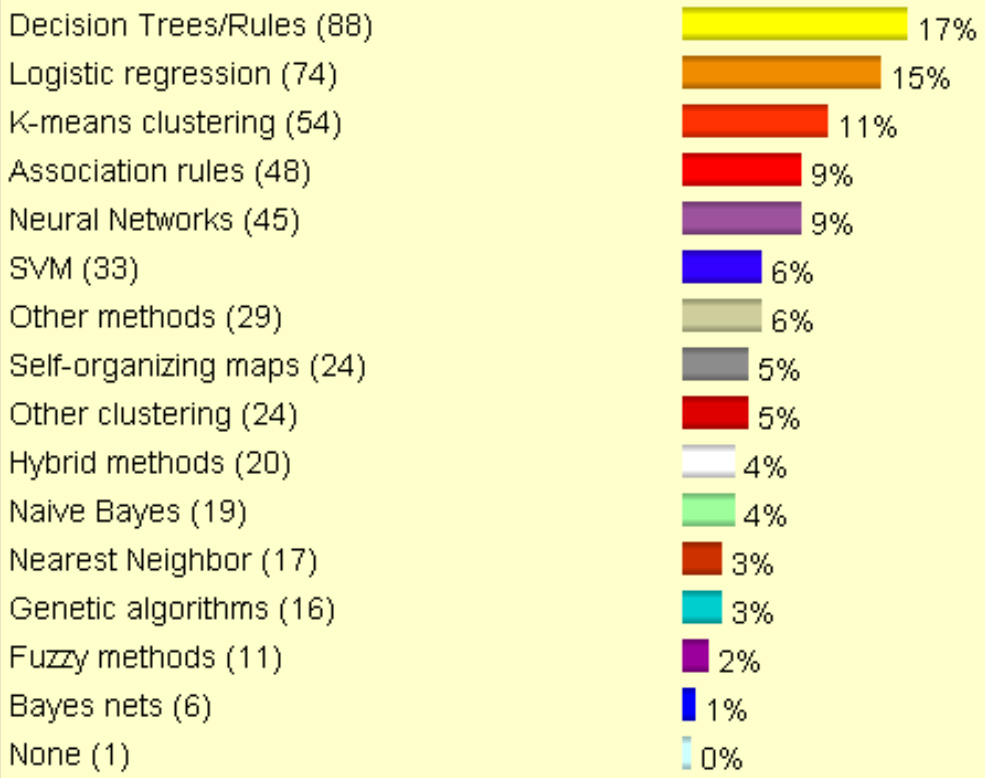
- phương pháp KNN (tên khác instance-based, lazy)
  - rất đơn giản, không có quá trình học
  - khi phân loại mất nhiều thời gian, do quá trình tìm kiếm k dữ liệu lân cận, sau đó phân loại dựa trên majority vote (hồi quy dựa trên giá trị trung bình)
  - kết quả phụ thuộc vài việc chọn khoảng cách sử dụng
  - có thể làm việc trên nhiều loại dữ liệu khác nhau
  - giải quyết các vấn đề về phân loại, hồi quy, gom nhóm, etc.
  - cho kết quả tốt, tuy nhiên độ phức tạp của quá trình phân loại khá lớn
  - được ứng dụng thành công trong hầu hết các lãnh vực tìm kiếm thông tin, nhận dạng, phân tích dữ liệu, etc.

Kỹ thuật DM thành công trong ứng dụng thực (2004)

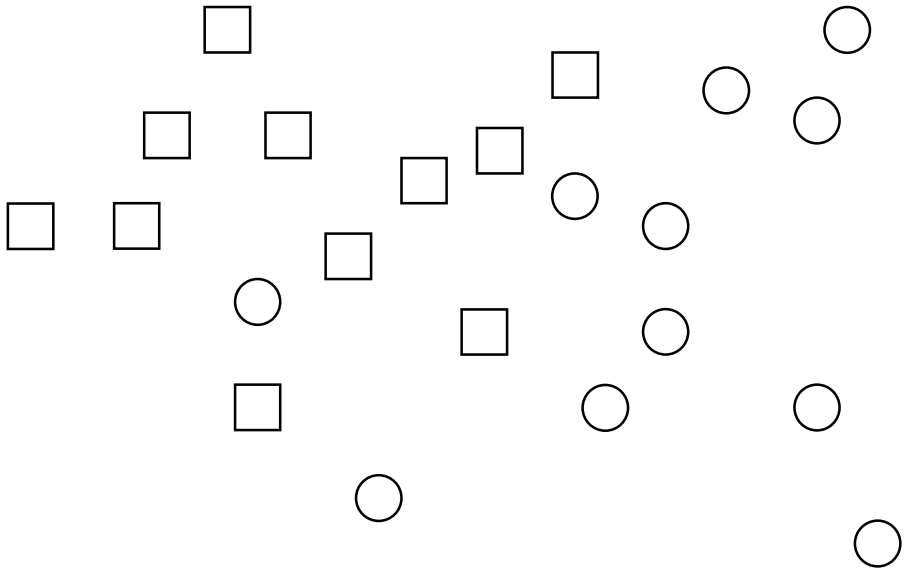
KDnuggets : **Polls** : Deployed data mining techniques

Poll

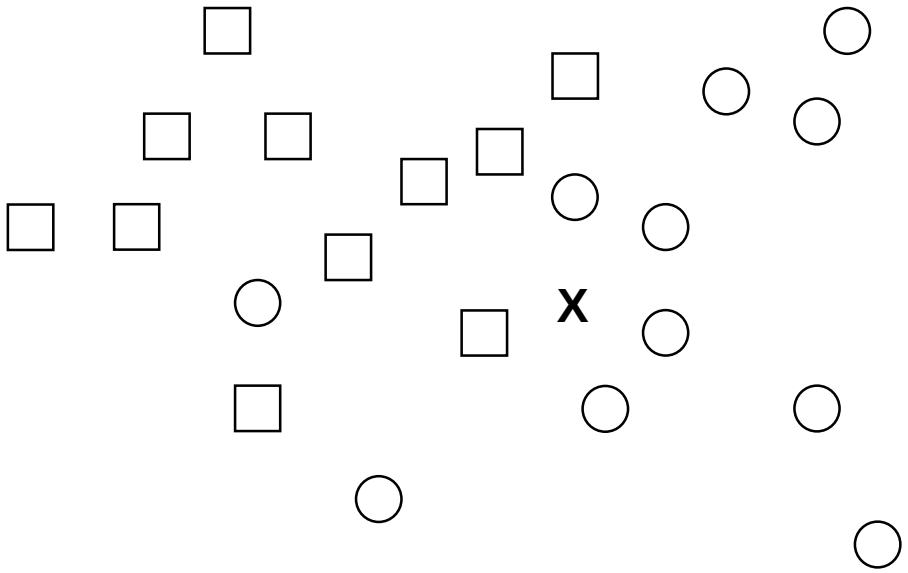
Which data mining techniques you used in a successfully deployed application? [173 voters, 509 votes total]



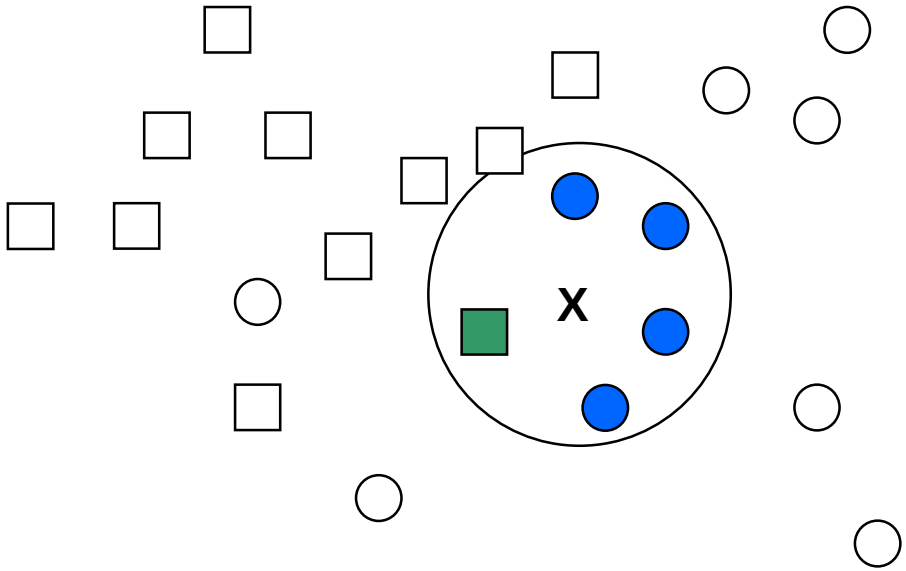
Phương pháp KNN



Phương pháp KNN



Phương pháp KNN



Phương pháp KNN

- khoảng cách *Minkowski*

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q)}$$

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ là 2 phần tử dữ liệu trong p-dimensional, $q >= 1$

- nếu $q = 1$, d là khoảng cách Manhattan

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

Phương pháp KNN

- *nếu q = 2, d là khoảng cách* Euclid

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

  - tính chất

    $d(i,j) \geq 0$

    $d(i,i) = 0$

    $d(i,j) = d(j,i)$

    $d(i,j) \leq d(i,k) + d(k,j)$

- *nên chuẩn hóa dữ liệu*

Phương pháp KNN

X1	X2	Lớp
0.45	5	?

X1	X2	Lớp
0.1	10	+1
0.2	25	+1
0.3	0	+1
0.5	11	-1
0.8	100	-1
0	50	+1
0.8	70	-1

D(Manhattan)
5.35
20.25
5.15
6.05
95.35
45.45
65.35

1NN ⇨ **lớp = +1**

Nhận xét

- *Thuộc tính X2 có miền giá trị 0..100) trong khi thuộc tính X1 có miền giá trị 0..1*
- *Kết quả phụ thuộc nhiều vào X2 (chênh lệch X2 lớn hơn so với X1)*
- *nên chuẩn hóa dữ liệu (chuẩn hóa thuộc tính X2 về giá trị 0..1*

new_val = (val – min)/(max – min)

Phương pháp KNN

X1	X2	Lớp
0.45	0.05	?

X1	X2	Lớp	D(Manhattan)
0.1	0.1	+1	0.4
0.2	0.25	+1	0.45
0.3	0	+1	0.2
0.5	0.11	-1	0.11
0.8	1	-1	1.3
0	0.5	+1	0.9
0.8	0.7	-1	1

1NN ⇨ **lớp = -1**

Nội dung

- Giới thiệu về KNN
- **Kết luận và hướng phát triển**

Phương pháp KNN

- thường rất chính xác, nhưng chậm do phải duyệt qua dữ liệu để tìm phần tử gần
- giả sử các thuộc tính có độ quan trọng như nhau
  - gán trọng số quan trọng cho mỗi thuộc tính
- chịu đựng được nhiễu
  - tham số k
  - xóa dữ liệu nhiễu (hơi khó ☹)
- thống kê đã sử dụng k-NN từ những năm 50s
  - khi dữ liệu lớn ($n \to \infty$) và $k/n \to 0$, lỗi gần với giá trị nhỏ nhất

Hướng phát triển

- tăng tốc cho quá trình tìm k phần tử lân cận
  - cấu trúc index
- chọn thuộc tính quan trọng
- gán trọng số cho các thuộc tính

Cám ơn !