Báo cáo kết quả bài toán dự đoán thuê bao ngủ đông (User activity level prediction)

Phòng Khoa học dữ liệu, TT VTTEK

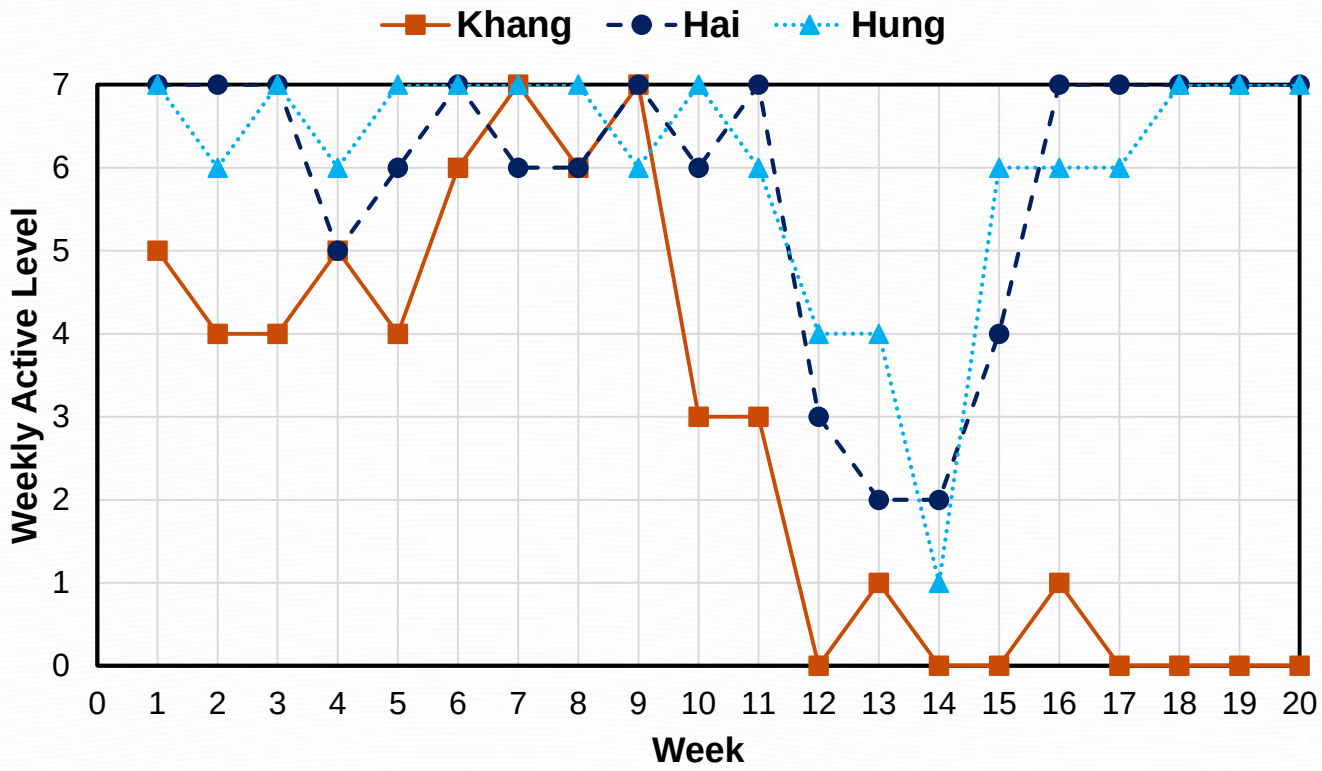
Hãy nói theo cách của bạn

Nội Dung

- Giới thiệu bài toán Dự đoán thuê bao ngủ đông
- Tóm tắt tiến độ thực hiện
- Mô hình dự đoán
- Kết quả thử nghiệm

Bài toán dự đoán thuê bao ngủ đông

- Mục tiêu: dự đoán trạng thái của thuê bao vào tuần tiếp theo là active hay inactive.



Thuê bao "Khang" inactive sau tuần thứ 11.
Mục tiêu: *tại tuần thứ 11, thuật toán đưa ra dự đoán thuê bao "Khang" sẽ inactive vào tuần tiếp theo*

Tiến độ thực hiện

Tiến độ thực hiện bài toán dự đoán thuê bao ngủ đông, và xây dựng hệ thống lưu trữ, xử lý dữ liệu lớn kết hợp Machine Learning

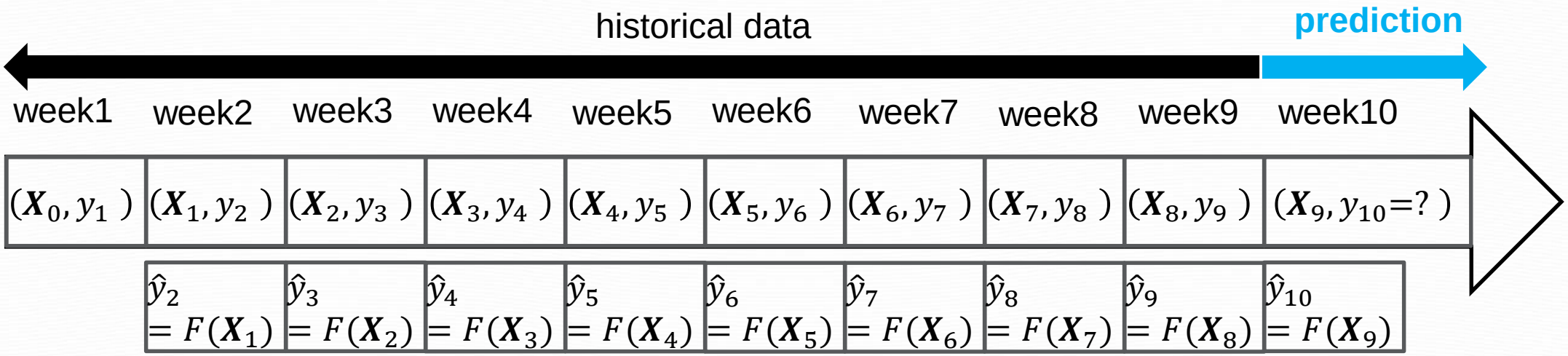
02/2018	03/2018	04/2018	05/2018	06/2018
Tìm hiểu CDR Đọc dữ liệu mẫu Call Detail Record, TT OCS, tìm hiểu các trường thông tin tiêu dung của thuê bao	**Nghiên cứu bài toán dự đoán thuê bao ngủ đông** Nghiên cứu một số bài báo, mô hình dự đoán thuê bao ngủ đông **Nghiên cứu hệ thống lưu trữ, xử lý dữ liệu lớn: Hadoop, Hbase**	**Xây dựng thuật toán dự đoán thuê bao ngủ đông** Xây dựng một thuật toán LR dự đoán thuê bao ngủ đông. **Hiển thị môt số thông tin tiêu dùng** **Xây dựng giải pháp lưu trữ, xử lý dữ liệu lớn: Hadoop, Hbase**	**Chuẩn bị dữ liệu để thử nghiệm thuật toán** Đọc dữ liệu CDR (200GB file bin) của 10 tuần, tách, lọc thông tin, xây dựng tập mẫu cho huấn luyện. **Setup Hadoop, Hbase trên cluster 6 server ảo**	**Thử nghiệm, đánh giá kết quả** Thử nghiệm thuật toán, thay đổi tham số, đánh giá kết quả. **Kết quả: precision 0.62, recall 0.79, F1 0.66.** **Ghi dữ liệu vào hệ thống cluster Hadoop, Hbase**

Xây dựng feature vector mô tả thuê bao

day	Data (MB)	Voice (minute)	SMS (#)	balance	recharge
1					
2					
3					
4					
5					
6					
7					
⋮					
72					

Mỗi thuê bao được mô tả bằng một mảng 2D

Xây dựng mô hình dự đoán thuê bao ngủ đông



Ý tưởng: đưa về bài toán Binary Classification sử dụng Logistic Regression. Dự đoán nhãn $y_{10} \in \{0,1\}$
đầu vào là tập dữ liệu gán dãn $(X_1, y_2), (X_2, y_3), \dots, (X_8, y_9)$

Model: tìm hàm F:

$$\min_F J = \sum_{t=2}^{T} \sum_{n=1}^{N} \left(y_t - F(X_{t-1})\right)^2$$

Logistic regression:

$$\min_{W_0, \{W_i\}_{n=1}^N} J = \sum_{t=2}^{T} \sum_{n=1}^{N} e^{-\alpha(T-t)}\, \ell\left(y_t^{(n)}, \left(W^{(0)} + W^{(n)}\right)^T X_t^{(n)}\right) + \gamma_0 \left|\left|W^{(0)}\right|\right|_2^2 + \gamma \sum_{n=1}^{N} \left|\left|W^{(n)}\right|\right|_2^2$$

Dữ liệu CDR

- Mô tả dữ liệu đầu vào
- Dữ liệu của bao nhiêu thuê bao?
  - Dữ liệu của 50000 users
- Dữ liệu ghi lại của ngày nào?
  - Trong khoảng thời gian 72 ngày (từ 01/03/2018 đến 12/05/2018), chia làm 10 tuần, trong đó 8 tuần cho training và 2 tuần cho test
- Các CDR nào được sử dụng?
  - CDR về tiêu dùng voice, sms, data
- Chuẩn hóa dữ liệu ra sao?
- Đơn vị tính các loại tiêu dùng của thuê bao?

Kết quả thử nghiệm

- Độ chính xác (Accuracy) = $\frac{Số\ thuê\ bao\ dự\ đoán\ chính\ xác}{Tổng\ số\ lượng\ người\ dùng}$
  - Accuracy = 0.62
- Tỉ lệ phát hiện inactive users (recall) = $\frac{\#\ inactive\ users\ được\ dự\ đoán\ đúng}{\#\ inactive\ users\ trên\ thực\ tế}$
  - Recall = 0.79
- Điểm F-beta = $\frac{5 * precision * recall}{4 * precision + recall}$ (ưu tiên recall hơn, mong muốn tìm tìm được càng nhiều inactive users trên thực tế càng tốt)
  - F-beta = 0.66
- So sánh với bài báo tham khảo (1) : gần như tương đương

(1) *Predicting User Activity Level in Social Networks*, Qiang Yang, Huawei Noah's Ark Lab, HongKong

Tìm hiểu hệ thống Recommender System Flytxt

- Mô tả input/output của hệ thống?
  - Data source
  - Data format
  - Product ID
  - ...
- Mô tả các tính năng của hệ thống?