1. GIỚI THIỆU THUẬT TOÁN DỰ ĐOÁN TRONG HỌC MÁY

1.1 Nội suy tuyến tính

Hồi qui tuyến tính đa biến là hồi qui tuyến tính với nhiều hơn một biến đầu vào. Hồi qui tuyến tính đa biến phổ biến hơn so với đơn biến vì trên thực tế rất hiếm các tác vụ dự báo chỉ gồm một biến đầu vào.

1.2 NỘI SUY SỬ DỤNG THUẬT TOÁN RANDOM FOREST

Random Forests (RF) là thuật toán có giám sát được sử dụng cho cả phân lớp và hồi quy. RF tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng. RF có nhiều ứng dụng, chẳng hạn như công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng…

RF được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting. Lý do chính là nó dùng trung bình của tất cả các dự đoán. Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy. RF cũng có thể xử lý các giá trị còn thiếu. Có hai cách để xử lý các giá trị này: sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu

RF là thuật toán cần nhiều thời gian để tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.

1.3 NỘI SUY SỬ DỤNG THUẬT TOÁN K-NEAREST NEIGHBORS

KNN (K-Nearest Neighbors) là một trong những thuật toán học có giám sát đơn giản nhất được sử dụng nhiều trong khai phá dữ liệu và học máy. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học (nên KNN được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán nhãn của dữ liệu mới.

2. Thực nghiệm

2.1 Đặc điểm khu vực thực nghiệm

Để triển khai thực nghiệm, chúng tôi sử dụng số liệu đo cao năm 2007 tại địa bàn thực nghiệm, bao gồm số liệu đo thủy chuẩn và số liệu đo GPS với tổng số 81 điểm phân bố tại 5 tuyến đo thủy chuẩn.

Với mục đích khảo sát độ chính xác đo cao GPS, chọn 7 điểm độ cao hạng III, phân bố dọc đường biên khu vực thực nghiệm. Từ số liệu đo thủy chuẩn và đo cao GPS đã tính được dị thường độ cao của 7 điểm nêu trên và coi các điểm đó là các điểm song trùng độ cao

Lớp	Ứng dụng	Lưu lượng	CoS	EXP	DSCP	PHB	Schedule	Queue	Congestion Avoidance
8	Dự phòng	Dự phòng	7	7	56	CS7	PQ	8	Tail Drop
7	Synchronization	PTP, NTP	6	6	48	CS6	PQ	7	Tail Drop
	Signaling	Sigtran, GTP Signal, IMS signaling VoLTE…							
	Network Protocol	OSPF, BGP, LDP, PIM							
	Stream Control	SCTP							
	Radio Network Control	FACH, RACH, PCH, MBMS, SRB							
	Conversational	HSPA/R99 Conversation	6	6	46	CS6			
3	4G, 5G data	4G, 5G Data	2	2	18, 20	AF21, AF22	WFQ	3	Tail Drop
1	Background	ADSL/FTTH truy cập dịch vụ Internet còn lại	0	0	0	BE	WFQ	1	WRED (Low: 60%, High: 100%)

Bảng 1: Tọa độ, độ cao hạng III là các điểm song trùng

Để khảo sát độ chính xác đo cao GPS, 74 điểm độ cao hạng IV phân bố trong khu vực thực nghiệm được sử dụng. Số liệu tọa độ, độ cao các điểm khảo sát được đưa ra bảng 2

Số TT	Tên điểm	Toạ độ phẳng		Độ cao (m)	
		x(m)	Trắc địa	Thuỷ chuẩn	Trắc địa
1	III-01	1201060.12	597671.84	5.473	4.866
2	III-02	1200959.05	593781.53	8.507	7.877
3	III-03	1204589.53	596719.49	2.122	1.501
4	III-04	1196868.26	594232.93	1.564	0.947
5	III-05	1201319.18	605534.10	10.454	9.887
6	III-06	1198744.19	603499.77	0.574	-0.001
7	III-07	1205274.12	602468.45	0.938	0.350
8	IV-01	1205298.75	602181.81	0.684	0.095

9	IV-02	1205171.47	601926.00	1.599	1.009
10	IV-03	1204979.05	601947.06	1.674	1.084
- - - ·	- - - ·	- - - ·	- - - ·	- - - ·	- - - ·
64	IV-57	1201723.68	594931.38	6.067	5.442
65	IV-58	1201605.30	594532.89	7.826	7.200
66	IV-59	1201536.10	594163.59	8.019	7.391
67	IV-60	1200972.22	594157.80	7.741	7.114
68	IV-61	1200841.20	594854.32	7.387	6.764
69	IV-62	1200614.81	595304.12	7.758	7.138
70	IV-63	1200436.83	595542.31	7.207	6.589
71	IV-64	1200200.40	595664.07	7.305	6.688
72	IV-65	1199843.75	595323.36	7.033	6.415
73	IV-66	1199430.49	595381.71	3.881	3.264

Bảng 2 : Tọa độ, độ cao các điểm độ cao hạng IV

2.2 KẾT QUẢ NỘI SUY ĐỘ CAO

Phương án 1 sử dụng 7 điểm độ cao hạng III để nội suy cho 67 điểm độ cao hạng IV

Bảng 3 : So sánh sai số trung phương (RMSE) và sai số tuyệt đối (MAE)đánh giá cho 67 điểm sử dụng ba thuật toán KNN, LR và RF

Sai số	**Thuật toán nội suy**		
	KNN (k=3) (m)	**LR (m)**	**RF (m)**
RMSE	2.15e-05	3.79e-06	3.26e-05
MAE	0.0034	0.0010	0.0044

Kết quả nội suy bằng LR tương tự như kết quả nội suy theo đa thức bậc 1 (Bảng 4). Điều này cho thấy thuật toán sử dụng trong báo cáo này là tin cậy.

Bảng 4 : Kết quả nội suy dị thường độ cao theo đa thức bậc 1 (kết quả trích dẫn từ luận văn Bùi Mai Khanh)

Số TT	**Tên điểm**	**Đo cao GPS (m)**			**Độ cao thủy chuẩn h_{TC},m**	**Độ lệch, m h_{GPS}-h_{TC}**
		H	**ζ**	**h_{GPS}=H- ζ**		
1	IV-01	0.095	-0.592	0.687	0.684	0.003
2	IV-02	1.009	-0.593	1.602	1.599	0.003

3	IV-03	1.084	-0.592	1.676	1.674	0.002
4	IV-04	0.449	-0.590	1.039	1.037	0.002
5	IV-05	-0.051	-0.591	0.540	0.538	0.002
6	IV-06	0.181	-0.591	0.772	0.770	0.002
7	IV-07	0.209	-0.590	0.799	0.796	0.003
8	IV-08	0.540	-0.588	1.128	1.127	0.001
9	IV-09	0.613	-0.590	1.203	1.202	0.001
10	IV-10	0.119	-0.589	0.708	0.707	0.001
11	IV-11	-0.128	-0.586	0.458	0.455	0.003
12	IV-12	0.039	-0.587	0.626	0.625	0.001
13	IV-13	0.066	-0.587	0.653	0.652	0.001
14	IV-14	-0.155	-0.586	0.431	0.429	0.002
15	IV-15	0.273	-0.586	0.859	0.858	0.001
16	IV-16	0.002	-0.584	0.586	0.586	0.000
17	IV-17	0.371	-0.584	0.955	0.954	0.001
18	IV-18	0.188	-0.583	0.771	0.770	0.001
19	IV-19	-0.427	-0.582	0.155	0.154	0.001
20	IV-20	0.640	-0.582	1.222	1.222	0.000
21	IV-21	0.608	-0.581	1.189	1.190	-0.001
22	IV-22	0.324	-0.581	0.905	0.906	-0.001
23	IV-23	0.702	-0.580	1.282	1.283	-0.001
24	IV-24	0.348	-0.578	0.926	0.927	-0.001
25	IV-25	0.867	-0.578	1.445	1.447	-0.002
26	IV-26	5.110	-0.607	5.717	5.717	0.000
27	IV-27	1.518	-0.603	2.121	2.120	0.001
28	IV-28	1.498	-0.600	2.098	2.097	0.001
29	IV-29	3.854	-0.597	4.451	4.450	0.001

30	IV-30	1.350	-0.592	1.942	1.942	0.000
31	IV-31	1.564	-0.587	2.151	2.151	0.000
32	IV-32	1.367	-0.578	1.945	1.945	0.000
33	IV-33	-0.758	-0.574	-0.184	-0.185	0.001
34	IV-34	-0.174	-0.573	0.399	0.398	0.001
35	IV-35	0.009	-0.572	0.581	0.581	0.000
36	IV-36	1.161	-0.616	1.777	1.777	0.000
37	IV-37	3.083	-0.614	3.697	3.697	0.000
38	IV-38	4.731	-0.612	5.343	5.344	-0.001
39	IV-39	5.415	-0.611	6.026	6.026	0.000
40	IV-40	8.265	-0.610	8.875	8.874	0.001
41	IV-41	9.613	-0.610	10.223	10.222	0.001
42	IV-42	8.864	-0.609	9.473	9.473	0.000
43	IV-43	7.619	-0.610	8.229	8.228	0.001
44	IV-44	8.235	-0.611	8.846	8.846	0.000
45	IV-45	7.821	-0.611	8.432	8.432	0.000
46	IV-46	5.613	-0.609	6.222	6.221	0.001
47	IV-47	7.121	-0.612	7.733	7.732	0.001
48	IV-48	6.662	-0.612	7.274	7.274	0.000
49	IV-49	4.107	-0.621	4.728	4.729	-0.001
50	IV-50	4.440	-0.621	5.061	5.062	-0.001
51	IV-51	6.616	-0.620	7.236	7.249	-0.013
52	IV-52	6.925	-0.619	7.544	7.545	-0.001
53	IV-53	6.958	-0.619	7.577	7.578	-0.001
54	IV-54	6.662	-0.620	7.282	7.283	-0.001
55	IV-55	6.367	-0.620	6.987	6.987	0.000
56	IV-56	5.456	-0.623	6.079	6.080	-0.001

57	IV-57	5.442	-0.624	6.066	6.067	-0.001
58	IV-58	7.200	-0.626	7.826	7.826	0.000
59	IV-59	7.391	-0.627	8.018	8.019	-0.001
60	IV-60	7.114	-0.627	7.741	7.741	0.000
61	IV-61	6.764	-0.623	7.387	7.387	0.000

Phương án 2:

Trong phương án này sử dụng 59 (80%) điểm độ cao để training và 15 (20%) điểm để testing

Bảng 5 : So sánh độ lệch giữa ba phương án KNN, LR và RF

Tên điểm	**Độ lệch (m)**		
	KNN (k=3)	**LR**	**RF**
IV-02	0.000680	0.00068	-0.001642
IV-09	0.001491	0.00137	-0.000527
IV-10	0.002092	0.00052	-0.000550
IV-24	-0.002429	-0.00261	0.001332
IV-25	0.000401	-0.00161	0.002743
IV-27	-0.001500	0.00065	-0.000497
IV-35	0.000588	-0.00012	0.001263
IV-39	0.000285	-0.00062	0.000507
IV-40	0.000000	-0.00008	-0.000690
IV-54	0.000069	0.00086	0.000680
IV-55	-0.001316	-0.00019	0.000130
IV-60	-0.001448	0.00009	0.000032
IV-61	0.001936	-0.00037	0.000026
IV-66	-0.000783	-0.00097	-0.000815
IV-51	0.012288	0.01239	0.012949
RMSE	1.1666e-05	1.2263e-05	1.12e-05
MAE	0.00182	0.001625	0.001542

Phương án 3:

Phương án này loại bỏ giá trị độ cao « bất thường » tại điểm IV-51 và thực hiện tương tự như phương án 2 ở trên. Sử dụng 59 điểm (80%) để training và 14 điểm (gần 20%) để testing

Bảng 6 : So sánh độ lệch giữa ba phương án KNN, LR và RF

Tên điểm	**Độ lệch (m)**		
	KNN (k=3)	**LR**	**RF**
IV-02	0.001	-0.002	0.001
IV-09	0.002	-0.001	0.001
IV-10	0.001	0.000	-0.000
IV-24	-0.002	0.001	-0.002
IV-25	-0.001	0.000	-0.001
IV-27	-0.001	-0.001	-0.000
IV-35	0.001	0.001	-0.000
IV-39	0.000	0.001	-0.000
IV-40	0.000	-0.001	-0.000
IV-54	0.001	0.001	0.001

IV-55	0.000	0.000	0.002
IV-60	-0.001	-0.000	0.000
IV-61	-0.001	-0.001	-0.001
IV-66	-0.001	-0.001	-0.001
RMSE	1.128e-06	5.784e-07	1.061e-06
MAE	0.00091	0.00061	0.00082