0254-6124/2022/42(2)-0199-07 *Chin. J. Space Sci.* 空间科学学报

WANG Zisiyu, SHI Liqin, LIU Siqing, ZHONG Qiuzhen, CHEN Yanhong, YAN Xiaohui, SHI Yurong, HE Xinran. *Kp* Index Prediction Based on Similarity Algorithm of Machine Learning (in Chinese). *Chinese Journal of Space Science*, 2022, **42**(2): 199–205. DOI:10.11728/cjss2022.02.210316030

基于机器学习相似度算法的 *Kp* 指数预报*

王子思禹[1,2,3] 师立勤[1,2,3] 刘四清[1,2,3] 钟秋珍[1,2,3]
陈艳红[1,3] 闫晓辉[1,2,3] 石育榕[1,2,3] 何欣燃[1,2,3]

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

3(中国科学院空间环境态势感知技术重点实验室 北京 100190)

摘 要 基于机器学习中的相似度算法，建立了在历史太阳风数据中寻找与当前太阳风特征相近事例的推荐模型，用来预报地磁 *Kp* 指数。使用 1998－2019 年间随机选择的 120 个太阳风事例作为测试数据集，该模型能够推荐得到历史上与输入太阳风造成相似地磁影响的太阳风事例，最优事例的 *Kp* 指数与实际值的均方根误差为 0.79，相关系数为 0.93。本文的推荐模型不仅能获得推荐的太阳风事例的地磁 *Kp* 指数用作预报，还可以给出太阳风特征参数按时间序列变化情况对比，让预报员可以更好地结合自身经验进行预报。

关键词 太阳风，机器学习，相似度算法，地磁 *Kp* 指数，预报

中图分类号 P353

Kp Index Prediction Based on Similarity Algorithm of Machine Learning

WANG Zisiyu[1,2,3] SHI Liqin[1,2,3] LIU Siqing[1,2,3] ZHONG Qiuzhen[1,2,3]
CHEN Yanhong[1,3] YAN Xiaohui[1,2,3] SHI Yurong[1,2,3] HE Xinran[1,2,3]

1(*National Space Science Center*, *Chinese Academy of Sciences*, *Beijing* 100190)

2(*University of Chinese Academy of Sciences*, *Beijing* 100049)

3(*Key Laboratory of Science and Technology on Environmental Space Situation Awareness*, *Chinese Academy of Sciences*, *Beijing* 100190)

Abstract The solar wind is the direct cause of the geomagnetic disturbance. In this paper, based on the feature selection and similarity algorithm of machine learning, a recommended model is established to search for cases whose characteristics are similar to the current solar wind in historical solar wind data, and to obtain the prediction of the geomagnetic *Kp* index. Tested on 120 solar wind cases

* 国家自然科学基金项目资助 (42074224)

2021-03-15 收到原稿，2021-12-19 收到修定稿

E-mail：wangzisiyu@bytedance.com

randomly selected from 1998 to 2019, the results show that the solar wind cases which have similar geomagnetic effects to the input solar wind can be worked out successfully by proposed model . And the root mean square error between the *Kp* index of the optimal case recommended by the model and the actual value is 0.79, and the correlation coefficient is 0.93. Different from traditional forecast models, the proposed recommended model in this paper can not only provide a geomagnetic *Kp* index as a forecast, but also give a clearer and more intuitive comparison of the changes between the solar wind characteristic parameters according to the time series. Even because the historical events have already happened, we can artificially find more dimensional information of the similar historical cases, which makes forecasters better combine their own experience in *Kp* index forecasting.

Key words Solar wind, Machine learning, Similarity algorithm, *Kp* index, Prediction

0 引言

太阳风是太阳高层大气向外流动所形成的超声速等离子体流。当由日冕物质抛射事件或冕洞产生的太阳风到达地球时，会向磁层注入更多能量和粒子，可能引起磁层扰动，从而引发地磁暴[1]。地磁暴会严重影响卫星等飞行器的性能和安全，从而对整个空间和地面的技术系统的正常运行造成威胁。因此，对太阳风造成的地磁扰动进行及时预报具有重要的实际意义。

随着机器学习技术的不断突破，其自适应、自学习的强大并行信息处理能力，在很多方面取得了突破性进展。特别是在目标推荐系统中的应用：通过分析用户历史行为，提取特征并进行相似度计算，可以解决用户从大量信息中挑选目标信息这样复杂且耗时的问题[2]。同时，在空间环境监测方面，作为衡量近地空间全球磁扰强度的重要指标之一的地磁 *Kp* 指数，已有长达 80 年无间断的数据[3]；太阳风方面也积累了几十年的卫星监测数据。因此，将机器学习技术应用于空间环境监测信息处理，可以更有效地进行地磁扰动预报。

目前已有的机器学习地磁 *Kp* 指数预报模型有：Rice 大学 Costello 模型[4]、Lund 天文台的 Boberg 模型[5]、USAF Wing *Kp* 模型(JHU/APL 模型)[6]、Rice 大学的 Bela 模型[7] 等，但这些方法都属于基于人工神经网络的回归问题，需要较多的模型参数实现地磁 *Kp* 指数的预报。同时，预报员也因为受限于机器学习的“黑盒”过程，不能充分结合自身经验判断进行预报。本文通过机器学习的聚类算法和相似度算法，建立了太阳风推荐模型。利用该模型可以计算并推荐出与输入太阳风事例相似的多个历史太阳风事例。通过对模型推荐的太阳风事例的后续 *Kp* 指数进行分析，以及推荐太阳风事例与输入的太阳风事例参数变化的对照分析，预报员可以更加有效地将自身经验融入到地磁扰动的预报当中。同时，根据历史相似太阳风事例的后续影响，预报员可以更早地对当前太阳风事例后续变化做出判断和预报。

1 数据

太阳风及行星际数据来源于 OMNI 网站*。OMNI 网站整合了来自 IMP 8、Geotail、Wind 和 Ace 卫星的太阳风数据。本文选取 ACE 卫星 1998－2019 年的太阳风数据，数据时频为 60 s。考虑到卫星太阳风数据可能存在缺失，对于缺失数据首先进行数据清洗：将个别的缺省数据近似取值为其前 60 s 对应的数据，保证了数据的完整性。OMNI 网站提供的太阳风参数包括太阳风速度 v、太阳风质子密度 N、太阳风温度 T、GSM 坐标下行星际磁场 B_y 和 B_z 分量的值以及行星际磁场强度 B。为了便于太阳风推荐模型的训练和学习，对太阳风参数数据均进行了归一化处理，消除不同量纲对计算的影响。

地磁 *Kp* 指数来源于 OMNI 网站整合自德国地球科学研究中心(GFZ)的结果。根据机器学习模型输入要求，将 *Kp* 指数进行临近值预处理：即 *Kp* 指数 3^0 转化为 3，3^-转化为 2.7，3^+转化为 3.3，以此类推。

* https://omniweb.gsfc.nasa.gov/

根据当前空间天气预报业务一般规范, 把 Kp 指数介于 5^- 和 6^+(包含 5^- 和 6^+)定义为小磁暴, 把 Kp 指数介于 7^- 和 9^0(包含 7^- 和 9^0)定义为大磁暴。

本文将世界时整点时刻后 3 h 内的太阳风数据作为一个太阳风事例, 进而将历史数据集划分成约 18 万个太阳风事例, 其中造成小磁暴影响的事例约 4800 个, 造成大磁暴影响的事例约 450 个。另外, 将时间上紧随其后的第一个 Kp 指数定义为与这一太阳风事例对应的 Kp 指数, 太阳风事例领先其对应的 Kp 指数 1～3 h, 即预报时间提前量为 1～3 h。

2 模型方法

太阳风是造成地磁暴的直接原因, 因此具有一致性特征的太阳风引起的地磁扰动也应该具有相似的特征。在历史数据足够多的条件下, 可以在历史数据中寻找到与当前太阳风特征相似的历史太阳风事例。历史太阳风事例引起的地磁变化, 就可以作为当前 Kp 指数预报的参考。

本文主要利用机器学习中的相似度算法来构建寻找相似特征太阳风事例的推荐模型。具体的步骤分为 3 步。第 1 步, 太阳风特征参数选取。可供选取的太阳风特征参数有速度、密度、温度、行星际磁场、行星际磁场 B_y 分量、行星际磁场 B_z 分量; 在相似度计算中, 过多的特征参数可能会导致推荐耗时倍增, 且推荐结果会被无影响或极低影响的参数干扰, 需要合理选取对地磁扰动起主要作用的参数作为相似度计算的特征参数。第 2 步, 参数权重计算。在地磁扰动中, 太阳风各种特征参数的影响作用不同, 需要对所选取的特征参数根据其对地磁扰动的影响计算相应的权重值。第 3 步, 推荐模型构建。本文以选定的太阳风特征参数作为输入, 基于机器学习中的相似度算法, 构建筛选相似太阳风事例的推荐模型。

2.1 太阳风特征参数选取

采用 XGBoost 算法模型进行太阳风特征参数选取。XGBoost[8] 是一个决策树训练模型, 该模型依据输入的特征参数在迭代建立决策树中的作用和价值, 即每个特征在所有树中作为划分属性的次数判断特征重要性, 继而通过每个属性分割点改进性能度量的量来计算单个决策树的重要性, 并由节点负责的观察数量加权, 最终将一个属性在所有提升树中的结果进行加权求和后然后平均, 得到重要性得分[9]。

以 1998－2019 年归一化后的太阳风速度 v、太阳风质子密度 N、太阳风温度 T、行星际磁场强度 B、行星际磁场 B_y、B_z 分量 1 h 的平均值作为输入, 以相应的 Kp 指数作为标签, 通过 XGBoost 模型迭代建立决策树, 对输入的六个特征参数进行了评分。评分高低即为 XGBoost 模型认为不同参数对地磁扰动(Kp 指数)影响的重要程度。评分结果如表 1。该结果也比较符合当前的研究结论, 即磁暴的主相发展与行星际磁场的南向分量和正值行星际电场密切相关[10]。为了避免过多特征参数造成推荐结果的混乱, 提高模型运行效率, 经过随机的部分样本的测试, 选择对地磁扰动影响最重要的两个参数: 太阳风风速 v 和行星际磁场 B_z 分量作为太阳风事例推荐模型的输入参数时, 推荐序列相似性和推荐耗时情况最好, 所以本文最终选择 v 和 B_z 作为相似度计算输入的特征。

2.2 参数权重计算

为了有效地计算筛选后的特征的权重, 避免多余特征的影响, 本文采用单层全连接神经网络对筛选后保留的太阳风风速 v, 行星际磁场 B_z 分量进行训练, 并评估它们各自对地磁扰动影响的权重。以归一化后的太阳风风速 v, 行星际磁场 B_z 分量作为单层神经网络的输入层参数, 时间分辨率是 1 h, 输出层为每小时相应的 Kp 指数, 损失函数为预测值和实际值的均方差函数, 优化器采用 Adam(Adaptive moment estimation)算法[11]。通过单层全连接神经网络训练后, 最终得到的权重比为 $|W_v|:|W_{Bz}| = 1.41:1$。

2.3 推荐模型构建

本文以机器学习中的相似度算法构建太阳风推荐模型, 主要采用了欧氏距离和动态时间规整(Dynamic Time Warping, DTW)两种算法。欧式距离是指 m 维空间中两个点之间的真实距离, 或者向量的自然长度, 在二维和三维空间中的欧氏距离就是两点之间的实际距离[12]。设空间中两点 X、Y 的坐标为 $X(X_1,X_2,\cdots,X_n), Y(Y_1,Y_2,\cdots,Y_n)$, 则 X 点与 Y 点间的欧式距离为

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2}. \tag{1}$$

表 1 XGBoost 对太阳风特征参数评分结果

Table 1 Scores of feature parameters using XGBoost

参数	v	B_z	N	B	T	B_y
评分	1647	850	794	724	474	400

当其输入为两组特征参量时，可以反映两组特征参量之间的综合差异。

动态时间规整算法(Dynamic Time Warping，简称 DTW 算法)是寻找时间序列相似的算法之一，其原理是给定两个序列 $Q=\{Q_1,Q_2,\cdots,Q_i,\cdots,Q_n\}$和 $C=\{C_1,C_2,\cdots,C_j,\cdots,C_m\}$，其长度分别是 n 和 m，构造一个 $n\times m$ 的矩阵网格，如图 1(a)所示，用红线和蓝线分别表示两个序列，以欧氏距离 D 为标准，矩阵元素 (i,j) 为 Q_i 和 C_j 两个点的距离 $D(Q_i,C_j)$。从序列起始段所在的矩阵角为边界条件，在满足连续性和单调性约束的同时，通过动态规划求得距离累积值最小的路径即为最佳路径[13]。

如图 1(b) 所示，动态时间规整算法认为，通过对时间序列进行局部非线性缩放，使两个序列形态尽可能对齐后的欧氏距离累计值才是两个时间序列的最小距离。

太阳风推荐模型包含初筛和精筛两个部分。初筛采用欧式距离方法，输入为一个太阳风事例中太阳风风速 v 与行星际磁场 B_z 的 1 h 平均值，3 h 即为维度 3×2 的矩阵。初筛通过逐个计算历史事例的 3×2 矩阵与输入事例对应的 3×2 矩阵的赋权欧式距离进行筛选，找到 3 个距离最近的历史事例，作为推荐事例。精筛采用动态时间规整算法，输入为一个太阳风事例中太阳风风速 v 和行星际磁场 B_z 的 60 s 值，3 h 即为维度 180×2 的矩阵。精筛计算历史上每个造成地磁暴的太阳风事例与输入事例之间的局部非线性缩放的欧氏距离，通过筛选找到最相似的 3 个历史太阳风事例作为推荐的太阳风事例。太阳风推荐模型实际运行时，首先对输入事例进行初筛，若初筛结果为输入太阳风事例不会造成地磁暴，则不继续进行更高精度的推荐，仅选用初筛的结果作为预报参考。若初筛结果认为可能发生磁暴，则以每分钟时间间隔的特征参数作为输入，继续通过精筛在引发磁暴的历史太阳风事例中寻找相似事例，得到参数时间序列变化相似的历史太阳风事例作为模型精筛的推荐结果。空间天气业务预报中，一般对可能发生磁暴的情况更为关注，因此推荐模型的精筛主要针对磁暴事例。推荐模型中没有全部采用精筛方式，也考虑到在大量历史数据中进行动态时间规整算法计算时，用时可能较长，降低预报实时性。

太阳风事例领先其对应的 Kp 指数 1～3 h，因此推荐模型的预报提前量为 1～3 h。

3 模型测试结果

以 1998－2019 年的太阳风数据作为历史太阳风数据对推荐模型进行测试。测试中，从 1998－2019 年历史太阳风事例中随机选择 120 个太阳风事例用作测试用例，这些测试事例覆盖了 1998－2019 年的全部年份，并且覆盖了从 0～9 全部 Kp 指数的等级，其中没有造成地磁暴影响的、造成小磁暴影响的和造成大磁暴影响的比例为 1:1:1，各 40 个。

选取测试用例的最优推荐事例(非自身的 top1)对应 Kp 指数与输入事例对应的实际 Kp 指数比较，测试结果如图 2。图中横轴是输入太阳风事例对应的实际地磁 Kp 指数，纵轴是推荐的历史太阳风事例的

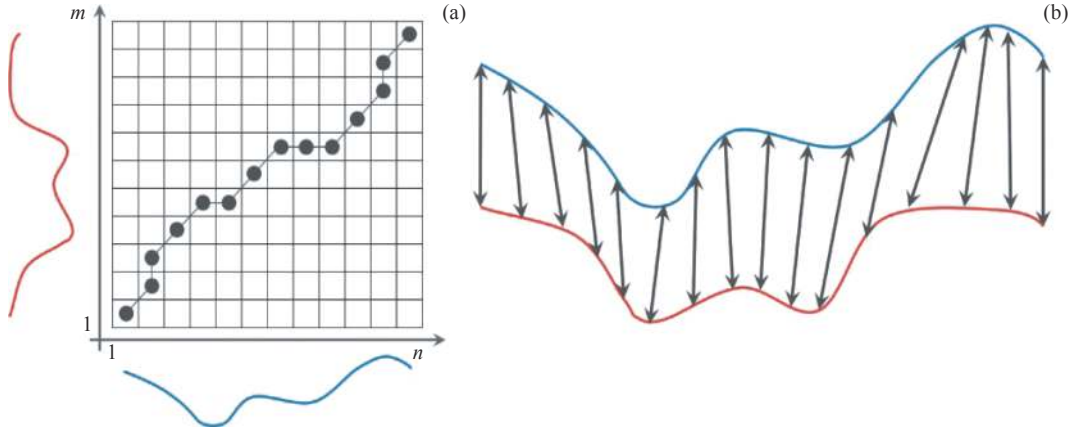

图 1 动态时间规整算法原理

Fig. 1 Principle of DTW

地磁 *Kp* 指数; I 区是无磁暴影响的太阳风, II 区是造成小磁暴影响的太阳风, III 区是造成大磁暴影响的太阳风。由图可知, 推荐太阳风事例的地磁 *Kp* 指数与输入太阳风事例的地磁 *Kp* 指数具有较好的一致性。*Kp* 平均绝对误差为 0.65, *Kp* 指数推荐值与实际值的均方根误差为 0.79, 相关系数为 0.93。

表 2 列出了典型的基于人工神经网络方法的已有 *Kp* 指数预报模型与本文模型的比较。对比结果表明, 本文的推荐模型优化了参数输入量的选取, *Kp* 预报结果的相关性更好, 预报结果误差较小。

本文模型可以推荐出 3 个最相似的事例, 测试中在 3 个最相似的事例中总能找出较为理想的推荐结果(推荐事例与实际 *Kp* 的偏差在 1 以内)。表 3 给出了推荐的三个最相似的事例的距离与 *Kp* 平均绝对误差(保留两位小数)的统计结果。由表 3 可知, 尽管存在一些距离与 *Kp* 指数最小变化精度不同导致的例外, 但距离越相近的事例, *Kp* 指数预报值越可能接近实测值。

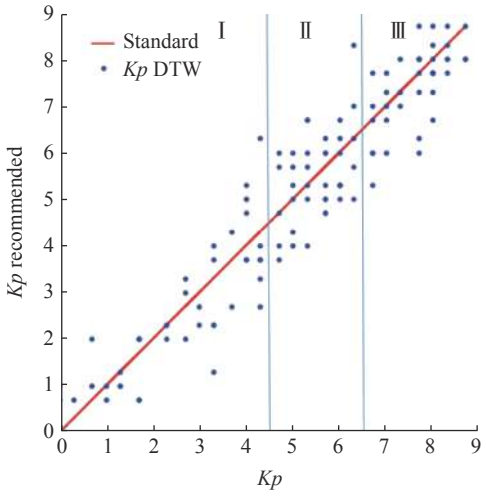

图 2 推荐模型测试结果

Fig. 2 Test results of the recommanded model

4 讨论

由于太阳风对地球磁层的能量输入方式极其复杂且高度动态, 进入地球磁层的能量可能快速耗散, 也可能暂时储存在磁层内部缓慢地释放, 所以现有各种模型很难给出精准预报。本文的推荐模型可以推荐出 3 个最相似事例, 预报员在实际应用中, 可以直接采用最优推荐结果, 也可以通过比对分析, 优选其中之一作为预报参考。推荐结果不光可以为预报员提供直接的 *Kp* 指数预报值, 还可以提供推荐事例与输入事例太阳风参数的时序变化对比, 供预报员结合自身经验进行预报调整。

图 3 中, 红线表示 2001 年第 90 天 03:00 UT 的输入太阳风参数变化, 绿线表示推荐事例之一的 2003 年第 324 天 11:00 UT 的太阳风参数变化, 其中两张子图纵轴分别表示太阳风风速 v 和行星际磁场 B_z 分量, 横轴均为太阳风事例的 3 h 的时间序列。由图可知, 输入的太阳风事例风速 v 更大, 根据人工预报经验, 当前输入的太阳风可能在地球磁层发生更多的磁重联, 单位时间内的磁通量可能更大, 因此输入到地球的能量可能更多, 则对推荐事例的结果可以向上微调。而实际结果正印证了本文的推测, 即实际输

表 2 本文模型与现有典型模型比较

Table 2 Comparison between the model proposed in this paper and the existing typical models

模型	输入量	提前时间/h	相关系数	均方根误差
Costello[4]	v, B, B_z	1	0.75	—
Boberg[5]	v, N, B_z	3	0.77	0.99
APL[6]	v, N, B, B_z, *Kp*	1	0.92	—
Bala Model 1[7]	Boyle index, *Kp*	1	0.863	0.71
Model 2	Boyle index, *Kp*	2	0.854	0.82
Model 3	Boyle index	1	0.852	1.12
Model 4	Boyle index	3	0.845	1.12
Liuyang Model[3]	v, N, B, B_y, B_z, $\mathrm{d}\phi/\mathrm{d}t$, $n^{1/2}v^2$	1～3.5	0.88	0.65
本文推荐模型	v, B_z	1～3	0.93	0.79

表 3 不同 *Kp* 指数下三个最相似事例的距离与平均 *Kp* 误差

Table 3 Distances of 3 recommanded cases and *Kp* error in different *Kp* indexes

Kp	测试事例个数	Top1(平均距离4.45) *Kp*平均绝对误差	Top2(平均距离5.08) *Kp*平均绝对误差	Top3(平均距离5.47) *Kp*平均绝对误差
0	3	0.70	1.43	1.23
1	11	0.46	0.55	0.71
2	5	0.38	0.24	0.24
3	9	0.81	0.43	0.48
4	12	0.78	1.81	1.33
5	20	0.66	1.12	1.36
6	20	0.90	0.73	0.85
7	13	0.51	0.52	0.57
8	24	0.57	1.13	0.98
9	3	0.23	0.88	1.68

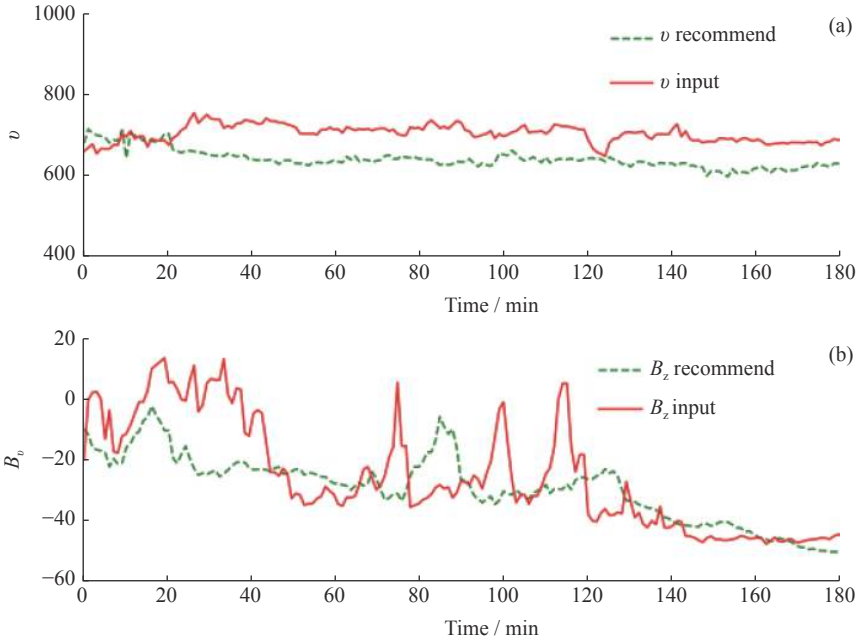

图 3 2001 年第 90 天 03:00 UT 太阳风与 2003 年第 324 天 11:00 UT 太阳风参数变化的对比

Fig. 3 Comparison of solar wind parameters at 03:00 UT on the 90th day in 2001 and 11:00 UT on the 324th day in 2003

入的太阳风事例对应相邻的地磁 *Kp* 指数为 9^-，推荐的太阳风事例对应相邻的地磁 *Kp* 指数为 8^-。

5 结论

本文以太阳风风速 v 和行星际磁场 B_z 作为输入特征参数，利用机器学习相似度算法计算输入事例的参数矩阵与历史事例的参数矩阵的距离，并根据距离排序结果进行推荐。与传统基于人工神经网络模型相比，本文的推荐模型基于 XGBoost 算法剔除了其它特征参数对地磁 *Kp* 指数影响的非线性关系，直接通过相似的太阳风历史事例的地磁扰动影响指导当前太阳风的预报，测试结果表明该模型具有较好的地磁 *Kp* 指数预报效果。

本文的推荐模型避免了人工神经网络模型预报的“黑盒”问题。预报员可以通过该模型获取 Kp 预测值,也可以从多个推荐结果中看到输入太阳风事例与推荐的历史事例的参数变化对比,使预报员可以在预报中更多地结合自身预报经验,优选调整预报结果。由于太阳风与磁层相互作用的复杂性,推荐模型只是通过历史相似事件给出当前输入太阳风的可能影响,Kp 指数预测精度的进一步提高也有赖于历史数据的积累。综上,本文的推荐模型可以为预报员提供一种更为实用的短时预报地磁暴的工具,也有利于预报员预报经验的提升。

参考文献

[1] KNIPP D J, MCQUADE M, KIRKPATRICK D. Understanding Space Weather and The Physics Behind it[M]. McGraw-Hill, 2011: 744

[2] LV Gang, ZHANG Wei. Survey of deep learning applied in recommendation system[J]. *Software Engineering*, 2020, **23**(2): 5-8 (吕刚, 张伟. 基于深度学习的推荐系统应用综述[J]. 软件工程, 2020, **23**(2): 5-8)

[3] LIU Yang, LUO Bingxian, LIU Siqing, *et al*. Kp forecast models based on neural networks[J]. *Manned Spaceflight*, 2013, **19**(2): 70-80 (刘杨, 罗冰显, 刘四清, 等. 基于神经网络方法的Kp预报模型[J]. 载人航天, 2013, **19**(2): 70-80)

[4] COSTELLO K A. Moving the Rice MSFM Into a Real-time Forecast Mode Using Solar Wind Driven Forecast Modules[D]. Houston: Rice University, 1998

[5] BOBERG F, WINTOFT P, LUNDSTEDT H. Real time Kp predictions from solar wind data using neural networks[J]. *Physics and Chemistry of the Earth, Part C:Solar, Terrestrial & Planetary Science*, 2000, **25**(4): 275-280

[6] WING S, JOHNSON J R, JEN J, *et al*. Kp forecast models[J]. *Journal of Geophysical Research*, 2005, **110**(A4): A04203

[7] BALA R, REIFF P H, LANDIVAR J E. Real-time prediction of magnetospheric activity using the Boyle index[J]. *Space Weather*, 2009, **7**(4): S04003

[8] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22 nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016

[9] LI Zhanshan, LIU Zhaogeng. Feature selection algorithm based on XGBoost[J]. *Journal on Communications*, 2019, **40**(10): 101-108 (李占山, 刘兆赓. 基于XGBoost的特征选择算法[J]. 通信学报, 2019, **40**(10): 101-108)

[10] TU Chuanyi, ZONG Qiugang, ZHOU Xuzhi. Solar-Terrestrial Space Physics[M]. 2 nd ed. Beijing: Science Press, 2020 (涂传诒, 宗秋刚, 周煦之. 日地空间物理学[M]. 2版. 北京: 科学出版社, 2020)

[11] KINGMA D P, BA J. ADAM: a method for stochastic optimization[C]//Proceedings of the 3 rd International Conference on Learning Representations. San Diego: ICLR, 2015

[12] VAN DER HEIJDEN F, DUIN R P, DE RIDDER D. Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB[M]. Hoboken: Wiley, 2004

[13] NI Qingqian, QIAO Jiyu, LIAN Zongkai. Improved K-means clustered pruning and DTW dynamic gesture recognition method[J]. *Modern Computer*, 2020(27): 20-25 (倪庆千, 乔冀瑜, 连宗凯. 改进K-means聚类剪枝的DTW动态手势识别方法[J]. 现代计算机, 2020(27): 20-25)

(责任编辑: 孙伟英)