Исследование значимости признаков переменных при локальном диагностировании полярных сияний / Ильясов Р.Р., Рахимов А.Р., Сахаутдинов А.А. и др.// Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 8 № 6(32) с. 109–114

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004

ИССЛЕДОВАНИЕ ЗНАЧИМОСТИ ПРИЗНАКОВ ПЕРЕМЕННЫХ ПРИ ЛОКАЛЬНОМ ДИАГНОСТИРОВАНИИ ПОЛЯРНЫХ СИЯНИЙ

[1]Ильясов Р.Р., Рахимов А.Р., Сахаутдинов А.А., Хайруллин Р.Д.
ФГБОУ ВО «Уфимский университет науки и технологий», Уфа, Россия (450076, г.Уфа, ул. Заки Валиди, 32), e-mail: [1]rustam.iljasov2015@yandex.ru.

Данная статья посвящена анализу временных рядов данных наблюдения полярных сияний на примере геофизической станции Ловозеро. Наиболее значимым является определение признаков, так или иначе влияющих на возникновение полярных сияний. В статье рассматриваются критерий Хи-квадрат, метод главных компонент, критерий Джини и метод прироста информации, использующихся для выявления наилучшим образом связанных переменных, при диагностировании исходных данных, а также модель, содержащая признаки и результаты наблюдения полярных сияний в зените.

Ключевые слова: полярные сияния, временной ряд, бинарная классификация, признаки, диагностирование.

INVESTIGATION OF THE SIGNIFICANCE OF VARIABLE SIGNS IN THE LOCAL DIAGNOSIS OF AURORAS

[1]Ilyasov R.R., Rakhimov A.R., Sakhautdinov A.A., Khairullin R.D.
Ufa University of Science and Technology, Ufa, Russia (450076, Ufa, Zaki Validi St., 32), e-mail: [1]rustam.iljasov2015@yandex.ru.

This article is devoted to analysis of time series of auroral observation data by the example of geophysical station Lovozero. The most important is to determine the features which somehow influence on the occurrence of auroras. The article considers Chi-square criterion, method of main components, Gini criterion and information increment method used to identify the best related variables, when diagnosing the initial data, as well as the model containing signs and results of aurora borealis observations in the zenith.

Keywords: auroras, time series, binary classification, attributes, diagnostics.

Определение значимости признаков в задаче бинарной классификации является важным шагом для оптимизации процесса моделирования и улучшения качества модели. Если модель использует множество признаков, которые не оказывают значимого влияния на результат классификации, то это может привести к переобучению модели и низкой точности предсказаний на новых данных. Использование нескольких методов может помочь в выявлении взаимосвязей между признаками.

Статистические методы

При определении значимости признаков в задаче бинарной классификации используются различные методы, наиболее распространёнными из которых являются корреляция Спирмена, критерий Джини, Mutual information, критерий Хи-квадрат и другие. В

данной статье рассмотрим конкретно критерий Хи-квадрат, метод главных компонент, критерий Джини и метод прироста информации. Критерий Хи-квадрат используется для проверки гипотезы о том, что две категориальные переменные независимы. Значение критерия Хи-квадрат показывает, насколько значима связь между признаком и классом. Чем выше значение критерия, тем сильнее связь между признаком и классом, и тем более значим признак. Однако, следует учитывать, что критерий Хи-квадрат может дать ложноположительные результаты, если используется небольшое количество признаков. Поэтому необходимо проводить дополнительные тесты и оценки модели, чтобы убедиться в правильности их выбора. Рассмотрим метод главных компонент. Он представляет собой ортогональное линейное преобразование, которое отображает данные из исходного пространства признаков в новое пространство меньшей размерности. С каждой главной компонентой связана определённая доля общей дисперсии исходного набора данных. В свою очередь, дисперсия, являющаяся мерой изменчивости данных, может отражать уровень их информативности. Также данные анализировались с помощью критерия Джини, используемого для оценки качества разделения классов в задачах классификации, в том числе бинарной. Он измеряет степень неоднородности (гетерогенности) данных в узле дерева решений или ветви разделения. При использовании алгоритма дерева решений для классификации данных, критерий Джини оценивает, насколько хорошо данный узел (или разделение) разделяет классы входных данных. Чем ниже значение критерия Джини, тем лучше разделение и тем более однородные классы получаются в результирующих поддеревьях. При выводе данных критерий интерпретировался так, что чем больше число получилось при выводе, тем сильнее зависимость признака. Ещё одним задействованным методом является метод прироста информации, основанный на понятии энтропии, которая измеряет неопределённость в данных. Данный метод использует метрику, называемую информационным коэффициентом, который измеряет разницу между начальной энтропией и суммарной энтропией после разбиения данных на основе признака. Чем больше прирост информации, тем более значимым является признак.

Оценка результатов исследования

Крайне важно, чтобы результаты исследования имели максимально достижимую точность. Для реализации этой цели был выполнен анализ данных, а так же ранжирование признаков по их значимости с применением языка программирования Python. Каждый из рассматриваемых признаков имеет отношение к полярным сияниям и, так или иначе, влияет на возникновение этого явления. Использование нескольких методов позволило увидеть более полную картину значимости признаков. На Рисунках 1-4 показаны результаты исследования.

feature	p_value
diff_LOZ_F	0.00e+00
diff_LOZ_Z	0.00e+00
diff_LOZ_E	0.00e+00
diff_LOZ_N	0.00e+00
diff_delta_LOZ_Z	0.00e+00
diff_delta_LOZ_E	0.00e+00
diff_LOZ_H	0.00e+00
diff_delta_LOZ_N	0.00e+00
LOZ_I	2.56e-305
LOZ_D	2.82e-256
AP	8.14e-154
abs_delta_LOZ_E	6.03e-78
abs_delta_LOZ_N	1.49e-75
abs_delta_LOZ_Z	8.30e-64
SMR	2.38e-34
delta_LOZ_N	7.37e-33
delta_LOZ_E	3.63e-32
SME	7.49e-28
delta_LOZ_Z	6.03e-27
LOZ_Z	1.16e-08
LOZ_E	4.58e-06
LOZ_N	3.73e-05
LOZ_H	1.32e-03
LOZ_F	5.66e-03

Рисунок 1 – Оценка результатов исследования критерием Хи-квадрат

```
diff_LOZ_N: 0.48657888351381356
delta_LOZ_N: 0.05785316656824976
LOZ_F: 0.056688710698743684
diff_LOZ_F: 0.05664553412785451
LOZ_D: 0.045565566227290376
LOZ_Z: 0.04413968388836338
LOZ_E: 0.03100976185713527
diff_LOZ_H: 0.020251529880875856
SMR: 0.019733642149983473
delta_LOZ_E: 0.0174463488331049
delta_LOZ_Z: 0.016768613128416827
diff_delta_LOZ_N: 0.01596559347732985
abs_delta_LOZ_E: 0.014999236369752087
LOZ_N: 0.014546000581291811
abs_delta_LOZ_N: 0.014492949380883356
diff_delta_LOZ_E: 0.013334266206174233
diff_LOZ_Z: 0.01100172768605685
AP: 0.010582696987286701
diff_delta_LOZ_Z: 0.010451649432754177
diff_LOZ_E: 0.010304137624196616
LOZ_H: 0.01004553537334826
abs_delta_LOZ_Z: 0.008962386354313555
SME: 0.006391669937802115
LOZ_I: 0.006240709714978787
```

Рисунок 2 – Оценка результатом исследования критерием Джини

```
diff_LOZ_E: 0.24916686344615310289
diff_delta_LOZ_E: 0.24915274236170426025
diff_LOZ_Z: 0.24675907709180067151
diff_delta_LOZ_Z: 0.24674910282598605527
diff_LOZ_F: 0.24666741356132521057
abs_delta_LOZ_N: 0.24622369880771904915
diff_delta_LOZ_N: 0.24193194351239896700
diff_LOZ_N: 0.24192606617217665699
diff_LOZ_H: 0.24168660620381290927
abs_delta_LOZ_E: 0.22759755799501368845
delta_LOZ_N: -0.22725851341734026878
SME: 0.22053153161021096795
AP: 0.22007294697885704249
LOZ_H: -0.21453731369515025018
abs_delta_LOZ_Z: 0.20771114049675770685
LOZ_N: -0.20635259368666289403
LOZ_I: 0.19078525104956406411
delta_LOZ_E: 0.19052126257368390272
SMR: -0.16665366974438378112
LOZ_D: 0.12340970300788797753
LOZ_E: 0.08093936307153734577
LOZ_F: -0.04239590690438941462
delta_LOZ_Z: -0.02373080534442671002
LOZ_Z: -0.01560708420469443178
```

Рисунок 3 – Оценка результатов методом главных компонент

```
diff_LOZ_F (0.315924)
diff_delta_LOZ_N (0.313000)
diff_LOZ_N (0.312154)
diff_LOZ_Z (0.310819)
diff_LOZ_H (0.308714)
diff_delta_LOZ_Z (0.308202)
abs_delta_LOZ_N (0.287647)
diff_LOZ_E (0.275748)
diff_delta_LOZ_E (0.275485)
delta_LOZ_N (0.261736)
SME (0.236958)
AP (0.236281)
abs_delta_LOZ_E (0.222483)
abs_delta_LOZ_Z (0.214448)
delta_LOZ_E (0.186486)
delta_LOZ_Z (0.180540)
SMR (0.150853)
LOZ_F (0.149657)
LOZ_Z (0.141910)
LOZ_H (0.138375)
LOZ_N (0.122769)
LOZ_D (0.121148)
LOZ_I (0.109257)
LOZ_E (0.102780)
```

Рисунок 4 – Оценка результатов исследования критерием прироста информации

Анализ данных о полярных сияниях показал, что наиболее важными признаками являются первые производные по северной (diff_LOZ_N), восточной (diff_LOZ_F) и вертикальной (diff_LOZ_Z) составляющими. Примерно на среднем уровне значимости находятся индексы геомагнитной активности (AP, SMR, SME), а менее важными являются простые северные (LOZ_N), восточные (LOZ_F) и вертикальные (LOZ_Z) составляющие.

Заключение

Таким образом, в ходе исследования были рассмотрены различные методы определения значимости признаков в задаче бинарной классификации. Анализ полученных результатов показал, что данные методы дают схожие результаты в определении значимости признаков. Естественно, результаты подвержены определённой погрешности, но комбинация методов позволяет учесть и устранить возможные искажения, связанные с этой погрешностью, а также помочь в качественном анализе и дальнейшем использовании полученных данных в исследовании полярных сияний.

Список литературы

1. А. В. Воробьев, В. А. Пилипенко, Т. А. Еникеев, Г. Р. Воробьева, О.И. Христодуло. Система динамической визуализации геомагнитных возмущений по данным наземных магнитных станций (2021). Научная визуализация 13.1: 162 - 176, DOI: 10.26583/sv.13.1.11
2. Воробьев А.В., Пилипенко В.А.. Подход к восстановлению гео- магнитных данных на базе концепции цифровых двойников. Солнеч- но-земная физика. 2021. Т. 7, No 2. С. 53–62. DOI: 10.12737/szf- 72202105.
3. Воробьев А.В., Пилипенко В.А., Сахаров Я.А., Селиванов В.Н. Статистические взаимосвязи вариаций геомагнитного поля, авро- рального электроджета и геоиндуцированных токов. Солнечно-земная физика. 2019. Т. 5, No 1. С. 48–58. DOI: 10.12737/szf-51201905.
4. Vorobev, A. V., V. A. Pilipenko, R. I. Krasnoperov, G. R. Vorobeva, and D. A. Lorentzen (2020), Short-term forecast of the auroral oval position on the basis of the “virtual globe” technology, Russ. J. Earth. Sci., 20, ES6001, doi:10.2205/2020ES000721.
5. Воробьев, А.В. Геоинформационная система для анализа динамики экстре- мальных геомагнитных возмущений по данным наблюдений наземных станций / А.В. Воробьев, В.А. Пилипенко, Т.А. Еникеев, Г.Р. Воробьева // Компьютерная оптика. – 2020. – Т. 44, No 5. – С. 782-790. – DOI: 10.18287/2412-6179-CO-707.

References

1. A.V. Vorobyov, V. A. Pilipenko, T. A. Enikeev, G. R. Vorobyova, O.I. Hristodulo. A system for dynamic visualization of geomagnetic disturbances based on data from ground-based magnetic stations (2021). Scientific visualization 13.1:162 - 176, DOI: 10.26583/sv.13.1.11
2. Vorobyov A.V., Pilipenko V.A. An approach to the restoration of geo-magnetic data based on the concept of digital twins. Solar-but-terrestrial physics. 2021. Vol. 7, No 2. pp. 53-62. DOI: 10.12737/szf- 72202105.

3. Vorobyev A.V., Pilipenko V.A., Sakharov Ya.A., Selivanov V.N. Statistical interrelations of variations of geomagnetic field, auroral electrojet and geoinduced currents. Solar-terrestrial physics. 2019. Vol. 5, No 1. pp. 48-58. DOI: 10.12737/szf-51201905.
4. Vorobev, A.V., V. A. Pilipenko, R. I. Krasnoperov, G. R. Vorobeva, and D. A. Lorentzen (2020), Short-term forecast of the auroral oval position on the basis of the “virtual globe" technology, Russ. J. Earth. Sci., 20, ES6001, doi:10.2205/2020ES000721.
5. Vorobyov, A.V. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances based on observations of ground stations / A.V. Vorobyov, V.A. Pilipenko, T.A. Enikeev, G.R. Vorobyova // Computer optics. – 2020. – Vol. 44, No. 5. – pp. 782-790. – DOI: 10.18287/2412-6179-CO-707.