# 1. Обзор литературы

Метод обзора: открыл google scholar и там взял самые цитируемые статьи, ссылающиеся на две оригинальные [15] [16] . Сейчас граница выставлена на 50 цитирований. Это 4 книги и около 150 статей. Статистика примерно следующая:

- 25 источников предлагают какие-либо усовершенствования,
- 103 источника либо используют оригинальный isoforest в своём пайплайне, либо сравнивают свои алгоритмы с ним, либо просто упоминают о его существовании,
- 19 источников осуществляют обзоры разных методов и подходов, упоминая или описывая в том числе изофорест,
- 7 источников не упоминают изофорест вообще,
- 2 источника найти не удалось.

## Обсуждаемые усовершенствования изоляционного леса

Здесь обсудим, какие усовершенствования предлагают различные авторы в доступной литературе.

### Статический лес

В статье [22] авторы предлагают проводить выборку пороговых значений для фич неравномерно от минимума до максимума, а по некоторому построенному на основе данных распределению. Распределение строится на основе линейной комбинации ядер, где ядра в центре имеют меньшую плотность вероятности, чем на краях. Таким образом вероятность выбрать из пороговое значение из плотного региона будет меньше, а разделение – лучше.

Статьи [27] [26] посвящены ”глубокому лесу”. Идея повторяет глубокие нейронные сети. Несколько раз каскадом повторяем одну и ту же операцию – строим лес, считаем скоры, приписываем скоры в дополнение к исходным фичам. Как именно происходит обучение я не догнал, но в целом подход хороший. Гиперпараметров значительно меньше, чем у нейронок.

Тут [21] в целом ничего интересного, кроме мысли о том, что изофорест можно применять в т.ч. и к категориальным данным. При этом сами они применяют его очень странно.

В статье [23] коллеги зачем-то выкидывают деревья из изофореста, но зато доказывают интересное наблюдение: средняя длина изолирующего пути не зависит от размерности для равномерно распределённых данных. Заодно выводят по-человечески ту самую формулу для средней длины пути.

Авторы [14], [25], [5] [6] применяют по сути изоляционный лес к географическим координатам машин, чтобы понять, когда машина уехала куда-то вообще не туда. У их подхода есть интересная особенность. Они не строят деревьев на траекториях, а прямо задают вопрос ”сколько делений выдержит конкретна эта трактория, если мы будем её изолировать от вот этого сабсемпла траекторий”.

Поступают они совершенно аналогично [23] , что намекает, что эта мысль у людей бродит. И правильно бродит – не всегда есть возможность нарезать данные в какой-то очевидной топологии. Например, если у нас данные состоят сплошь из категориальных

данных. Как делить тогда? Или если данные разреженные – тот же вопрос. Не всегда вообще в данных присутствуют все измерения. Такой подход в принципе может упростить работу с инженерией фич.

В статье [3] авторы создают гибрид между LOF и изофорестом. Выглядит любопытно. Но это уже совсем не деревья получаются. Никакой вероятностной интерпретацией и не пахнет. Да и во много опирается на понятие расстояния, от которого изофорест удачно дистанцировался.

Интересная попытка обогатить изофорест – это extended isolation forest [11] . Вместо выбора случайной фичи выбирается случайное направление. Это хорошая заявка на успех, но почему-то по тестам в той же статье получается не суперархишикарно.

Считаю, что причина в том, что идея не доведена до логического завершения. Посмотрел код – у них вместо порогового значения скалярного произведения зачем-то выбирается случайная точка из гиперкуба, ограничивающего данные. Явно есть возможность нахалявку доработать и получить результат наверняка более впечатляющий. Либо узнать нечто новое про изоляционный лес.

Чёрт, это уже сделано в [13] . Буквально год назад.

В статье [24] приводится подход, где куча всего изменено, а доказательной базы никакой. Приводится как раз идея оценивать собственно сами распределения, а не матожидание логарифма. Даже какая-то теория прилеплена, но не видно доказательств, что это лучше обычного изофореста.

**Кластеры аномалий**

Много кто заинтересован в поиске целых кластеров аномалий. Глубоко в эти статьи ещё не вчитывался.

Авторы оригинальной статьи предлагали накрутить поиск аномальных кластеров в [17] .

Что именно сделали в статье [12] я искренне не понял. Они что-то как-то порезали на кластеры. Как – хрен знает. Вчитываться в детали пока не собираюсь, т.к. направление мыслей не впечатляет.

**Стриминг**

Про стриминг вообще чёртовы наркоманы пишут. Либо они прикидываются.

Авторы [8] предлагают резать входные данные на блоки и на основе этих блоков строить новые леса. Написано достаточно мутно, поэтому не ясно ни то, как они считают скоры блоков, ни то, при каких обстоятельствах они считают, что нашли аномалию. Очень скользкие типы.

Очередная статья про стриминг [18] мутная. Режут данные на разные виды аномальностей – подальше от регулярных данных и поближе, покрывают шарами те, которые поближе, доращивают деревья. Какая-то откровенная дичь, которую я вообще не затащил.

Ещё про стриминг – [10] . Здесь уже интереснее. Сперва авторы меняют алгоритм построения дерева – зачем-то вешают разные веса на разные размерности в зависимости от диаметра данных на этой размерности. Таким образом вновь убивая независимость изофореста от определения расстояния. Но зато у них какие-то свои прикольные мысли на тему того, как считать скор аномалий и как жить в стриме.

Ещё про стриминг – [24] . Та самая статья, обсуждающая то, как оценивать именно плотность вероятности. В названии сказано про стриминг, но конкретно стриминга внутри там очень немного.

**Активное обучение**

В статье [19] предлагают попробовать интересный вид уменьшения размерностей – модернизировать PCA по SVD-разложению минимаксом. Если верить их графикам, штука очень даже хорошо рабочая.

Что по лесу, то авторы делают важное наблюдение, что при наличии парочки размеченных аномалий уже можно собрать статистики по тому, какие фичи позволяют эти аномалии дискриминировать (аналогичную мысль повторяют авторы статьи [9] ).

Отдельно интересно, что эти коллеги фитят F-распределение к скорам аномальности семплов на изоляционном лесе. Это какой-то теоретический результат?

Статьи [20] и [7] обсуждаются здесь.

**Общеметодическое**

Отдельно от всех упомяну статью [4] , которая обсуждает, как вообще измерять и сравнивать перморм алгоритмов поиска аномалий. Статья не впечатляющая на первый взгляд, но возможно что-то полезное для наших дел можно оттуда подчерпнуть.

Другая статья для отдельного упоминания – [2] . Авторы обсуждают вопросы bias-variance tradeoff, bagging и subsampling в применении к поиску аномалий.

Ещё отдельно интересно почитать повнимательнее книгу [1] . Там хоть всё по поверхности, но зато ссылок много, к которым следует приглянуться.

## Список источников


1. *Aggarwal C. C.*, *Sathe S.* Outlier Ensembles. — Cham : Springer International Publishing, 2017. — ISBN 978-3-319-54764-0 978-3-319-54765-7. — DOI: 10.1007/978-3-319-54765-7. — URL: http://link.springer.com/10.1007/978-3-319-54765-7 (дата обр. 19.07.2022).
2. *Aggarwal C. C.*, *Sathe S.* Theoretical Foundations and Algorithms for Outlier Ensembles // ACM SIGKDD Explorations Newsletter. — 2015. — 29 сент. — Т. 17, № 1. — С. 24—47. — ISSN 1931-0145. — DOI: 10.1145/2830544.2830549. — URL: https://doi.org/10.1145/2830544.2830549 (дата обр. 19.07.2022).
3. *Bandaragoda T. R.* [и др.]. Efficient Anomaly Detection by Isolation Using Nearest Neighbour Ensemble // 2014 IEEE International Conference on Data Mining Workshop (2014 IEEE International Conference on Data Mining Workshop (ICDMW)). — Shenzhen, China : IEEE, 12.2014. — С. 698—705. — ISBN 978-1-4799-4274-9 978-1-4799-4275-6. — DOI: 10.1109/ICDMW.2014.70. — URL: http://ieeexplore.ieee.org/document/7022664/ (дата обр. 20.07.2022).
4. *Campos G. O.* [и др.]. On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study // Data Mining and Knowledge Discovery. — 2016. — 1 июля. — Т. 30, № 4. — С. 891—927. — ISSN 1573-756X. — DOI: 10.1007/s10618-015-0444-8. — URL: https://doi.org/10.1007/s10618-015-0444-8 (дата обр. 19.07.2022).
5. *Chen C.* [и др.]. iBOAT: Isolation-Based Online Anomalous Trajectory Detection // IEEE Transactions on Intelligent Transportation Systems. — 2013. — Июнь. — Т. 14, № 2. — С. 806—818. — ISSN 1558-0016. — DOI: 10.1109/TITS.2013.2238531.

6. *Chen C.* [и др.]. Real-Time Detection of Anomalous Taxi Trajectories from GPS Traces // Mobile and Ubiquitous Systems: Computing, Networking, and Services / под ред. A. Puiatti, T. Gu. — Berlin, Heidelberg : Springer, 2012. — С. 63—74. — ISBN 978-3-642-30973-1. — DOI: 10.1007/978-3-642-30973-1_6.
7. *Das S.* [и др.]. Incorporating Expert Feedback into Active Anomaly Discovery. —.
8. *Ding Z.*, *Fei M.* An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data Using Sliding Window // IFAC Proceedings Volumes. — 2013. — 1 янв. — Т. 46, № 20. — С. 12—17. — (3rd IFAC Conference on Intelligent Control and Automation Science ICONS 2013). — ISSN 1474-6670. — DOI: 10.3182/20130902-3-CN-3020.00044. — URL: https://www.sciencedirect.com/science/article/pii/S1474667016314999 (дата обр. 19.07.2022).
9. *Gavai G.* [и др.]. Detecting Insider Threat from Enterprise Social and Online Activity Data // Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats. — New York, NY, USA : Association for Computing Machinery, 16.10.2015. — С. 13—20. — (MIST '15). — ISBN 978-1-4503-3824-0. — DOI: 10.1145/2808783.2808784. — URL: https://doi.org/10.1145/2808783.2808784 (дата обр. 19.07.2022).
10. *Guha S.* [и др.]. Robust Random Cut Forest Based Anomaly Detection on Streams // Proceedings of The 33rd International Conference on Machine Learning (International Conference on Machine Learning). — PMLR, 11.06.2016. — С. 2712—2721. — URL: https://proceedings.mlr.press/v48/guha16.html (дата обр. 19.07.2022).
11. *Hariri S.*, *Kind M. C.*, *Brunner R. J.* Extended Isolation Forest // IEEE Transactions on Knowledge and Data Engineering. — 2021. — Апр. — Т. 33, № 4. — С. 1479—1489. — ISSN 1558-2191. — DOI: 10.1109/TKDE.2019.2947676.
12. *Karczmarek P.* [и др.]. K-Means-based Isolation Forest // Knowledge-Based Systems. — 2020. — 11 мая. — Т. 195. — С. 105659. — ISSN 0950-7051. — DOI: 10.1016/j.knosys.2020.105659. — URL: https://www.sciencedirect.com/science/article/pii/S0950705120301064 (дата обр. 20.07.2022).
13. *Lesouple J.* [и др.]. Generalized Isolation Forest for Anomaly Detection // Pattern Recognition Letters. — 2021. — Сент. — Т. 149. — С. 109—119. — ISSN 01678655. — DOI: 10.1016/j.patrec.2021.05.022. — URL: https://linkinghub.elsevier.com/retrieve/pii/S0167865521002063 (дата обр. 19.07.2022).
14. *Lin Q.* [и др.]. Disorientation Detection by Mining GPS Trajectories for Cognitively-Impaired Elders // Pervasive and Mobile Computing. — 2015. — Май. — Т. 19. — С. 71—85. — ISSN 15741192. — DOI: 10.1016/j.pmcj.2014.01.003. — URL: https://linkinghub.elsevier.com/retrieve/pii/S1574119214000200 (дата обр. 20.07.2022).
15. *Liu F. T.*, *Ting K. M.*, *Zhou Z.-H.* Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining (2008 Eighth IEEE International Conference on Data Mining (ICDM)). — Pisa, Italy : IEEE, 12.2008. — С. 413—422. — ISBN 978-0-7695-3502-9. — DOI: 10.1109/ICDM.2008.17. — URL: http://ieeexplore.ieee.org/document/4781136/ (дата обр. 16.07.2021).

16. *Liu F. T.*, *Ting K. M.*, *Zhou Z.-H.* Isolation-Based Anomaly Detection // ACM Transactions on Knowledge Discovery from Data. — 2012. — Март. — Т. 6, № 1. — С. 1—39. — ISSN 1556-4681, 1556-472X. — DOI: 10.1145/2133360.2133363. — URL: https://dl.acm.org/doi/10.1145/2133360.2133363 (дата обр. 16.07.2021).

17. *Liu F. T.*, *Ting K. M.*, *Zhou Z.-H.* On Detecting Clustered Anomalies Using SCiForest // Machine Learning and Knowledge Discovery in Databases / под ред. J. L. Balcázar [и др.]. — Berlin, Heidelberg : Springer, 2010. — С. 274—290. — ISBN 978-3-642-15883-4. — DOI: 10.1007/978-3-642-15883-4_18.

18. *Mu X.*, *Ting K. M.*, *Zhou Z.-H.* Classification Under Streaming Emerging New Classes: A Solution Using Completely-Random Trees // IEEE Transactions on Knowledge and Data Engineering. — 2017. — Авг. — Т. 29, № 8. — С. 1605—1618. — ISSN 1558-2191. — DOI: 10.1109/TKDE.2017.2691702.

19. *Puggini L.*, *McLoone S.* An Enhanced Variable Selection and Isolation Forest Based Methodology for Anomaly Detection with OES Data // Engineering Applications of Artificial Intelligence. — 2018. — 1 янв. — Т. 67. — С. 126—135. — ISSN 0952-1976. — DOI: 10.1016/j.engappai.2017.09.021. — URL: https://www.sciencedirect.com/science/article/pii/S095219761730235X (дата обр. 20.07.2022).

20. *Siddiqui M. A.* [и др.]. Feedback-Guided Anomaly Discovery via Online Optimization // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. — New York, NY, USA : Association for Computing Machinery, 19.07.2018. — С. 2200—2209. — (KDD '18). — ISBN 978-1-4503-5552-0. — DOI: 10.1145/3219819.3220083. — URL: https://doi.org/10.1145/3219819.3220083 (дата обр. 19.07.2022).

21. *Sun L.* [и др.]. Detecting Anomalous User Behavior Using an Extended Isolation Forest Algorithm: An Enterprise Case Study. — 21.09.2016. — URL: http://arxiv.org/abs/1609.06676 (дата обр. 20.07.2022).

22. *Tokovarov M.*, *Karczmarek P.* A Probabilistic Generalization of Isolation Forest // Information Sciences. — 2022. — 1 янв. — Т. 584. — С. 433—449. — ISSN 0020-0255. — DOI: 10.1016/j.ins.2021.10.075. — URL: https://www.sciencedirect.com/science/article/pii/S0020025521010999 (дата обр. 19.07.2022).

23. *Vinh N. X.* [и др.]. Discovering Outlying Aspects in Large Datasets // Data Mining and Knowledge Discovery. — 2016. — 1 нояб. — Т. 30, № 6. — С. 1520—1555. — ISSN 1573-756X. — DOI: 10.1007/s10618-016-0453-2. — URL: https://doi.org/10.1007/s10618-016-0453-2 (дата обр. 20.07.2022).

24. *Wu K.* [и др.]. RS-Forest: A Rapid Density Estimator for Streaming Anomaly Detection // 2014 IEEE International Conference on Data Mining (2014 IEEE International Conference on Data Mining). — 12.2014. — С. 600—609. — DOI: 10.1109/ICDM.2014.45.

25. *Zhang D.* [и др.]. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces // Proceedings of the 13th International Conference on Ubiquitous Computing - UbiComp '11 (The 13th International Conference). — Beijing, China : ACM Press, 2011. — С. 99. — ISBN 978-1-4503-0630-0. — DOI: 10.1145/2030112.2030127. — URL: http://dl.acm.org/citation.cfm?doid=2030112.2030127 (дата обр. 19.07.2022).

26. *Zhou Z.-H.*, *Feng J.* Deep Forest // National Science Review. — 2019. — 1 янв. — Т. 6, № 1. — С. 74—86. — ISSN 2095-5138. — DOI: 10.1093/nsr/nwy108. — URL: https://academic.oup.com/nsr/article/6/1/74/5123737 (дата обр. 19.07.2022).

27. *Zhou Z.-H.*, *Feng J.* Deep Forest: Towards An Alternative to Deep Neural Networks // Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (Twenty-Sixth International Joint Conference on Artificial Intelligence). — Melbourne, Australia : International Joint Conferences on Artificial Intelligence Organization, 08.2017. — С. 3553—3559. — ISBN 978-0-9992411-0-3. — DOI: 10.24963/ijcai.2017/497. — URL: https://www.ijcai.org/proceedings/2017/497 (дата обр. 19.07.2022).