УДК 004.85

А.А. Шолохова

ПОИСК АНОМАЛИЙ В СЕНСОРНЫХ ДАННЫХ НА ПРИМЕРЕ АНАЛИЗА ДВИЖЕНИЯ МОРСКОГО СУДНА

Санкт-Петербургский государственный университет

В статье рассмотрены алгоритмы поиска аномалий в сенсорных данных применительно к задаче анализа движения морских судов. Современное судно снабжено большим количеством датчиков, непрерывно записывающих показатели функционирования различных его подсистем. В случае, когда организован сбор и хранение такой информации, открываются возможности применения интеллектуальных систем анализа данных. Данные задачи относятся как к вопросам обеспечения безопасности движения (анализ и предупреждение опасных маневров), так и к проблемам повышения экономической эффективности (повышенное потребление топлива) для судовладельцев. Датчики, установленные на судне, могут генерировать числовые данные с частотой от десяти секунд для телеметрии до одной минуты для навигационных параметров. Учитывая такой большой объем информации, становится очевидной необходимость развития автоматических систем поддержки принятия решений. Прикладными областями подобных систем могут служить, например, задачи предупреждения опасных маневров, прогнозирования поломок, предотвращения столкновений, оптимизации потребления топлива. В статье рассмотрено применение алгоритмов обучения без учителя для анализа навигационных данных (координаты судна, его скорость, курс, глубина и пр.), а также пример прогнозирования потребления топлива на основе регрессионных моделей. Приведены описания различных математических подходов и демонстрация их применения на реальных данных. В заключении рассмотрены варианты дальнейшего развития и усложнения рассмотренных методов.

Ключевые слова: поиск аномалий, сенсорные данные, опасные маневры, прогнозирование потребления топлива

Введение. В настоящее время морская область переходит к этапу цифровой трансформации [1]. Появляются и внедряются системы интернета вещей (IoT), больших данных (Big Data), машинного обучения. Внедрение подобных автоматических систем поддержки принятия решения в основном направлено на две группы задач: к первой относятся вопросы повышения безопасности судовождения, анализа и предупреждения опасных маневров, обнаружение нештатных режимов движения судна; ко второй группе следует отнести экономические эффекты от оптимизации режимов потребления топлива, прогнозирование возможных поломок.

Решение указанных задач требует, как развития инфраструктуры по сбору и хранению данных, так и разработки новых подходов анализа данных. Традиционно в задаче поиска аномалий выделяют несколько подходов, различающихся применяемыми математическими методами.

К первой группе методов можно отнести подходы, основанные на статистических тестах. Также подобные алгоритмы применяются в анализе сенсорных данных при моделировании их временными рядами. Такой подход применен, например, в работе [2], где сенсорные данные моделируются авторегрессионной моделью, а для обнаружения аномалий используется теория проверки гипотез.

Другую группу составляют метрические и итерационные алгоритмы обнаружения выбросов. Простейшим примером такого подхода является построение выпуклой оболочки в многомерном пространстве в предположении, что выбросы в данных будут лежать на ней. В качестве другого варианта можно привести метод k ближайших соседей, в котором применительно к аномалиям предполагается, что выбросы имеют “большое” расстояние до соседних точек пространства. Применение подобных алгоритмов для задач поиска аномалий рассмотрено в работах [3, 4]. К недостаткам таких подходов относятся вычислительная трудоемкость и сложность выбора подходящей метрики. Причем если производительность метода можно повысить при помощи современных технологий параллельных и распределенных вычислений, то выбор метрики, удачно отражающей структуру многомерных данных, остается открытым вопросом и требует отдельного рассмотрения. Более подробный обзор указанных подходов может быть найден, например, в работах [5, 6].

Модельные тесты обнаружения аномалий основаны на построении некоторой модели данных. Точки пространства, которые существенно отклоняются от модели, можно считать аномалиями. Преимуществами таких подходов являются возможность учета предметной области и задание требуемого функционала качества [7, 8].

С точки зрения методов машинного обучения задача нахождения аномалий в данных может быть выделена в отдельный класс. К самым популярным алгоритмам относятся метод опорных векторов с одним классом и алгоритм изолирующего леса [9]. Основная идея первого подхода заключается в применении классического метода опорных векторов для отделения точек пространства от начала координат. Второй подход основан на построении разделяющих деревьев, описывающих данные. Деревья строятся, пока каждая точка рассматриваемого пространства не окажется в листовом узле, после чего критерием нормальности может служить глубина (или среднее значение глубин для леса) этого узла. К достоинствам данных алгоритмов относится их простая параметризуемость, возможность задавать различные метрики и количество степеней свободы модели. Также в процессе работы алгоритмов строится разделяющая гиперплоскость, расстояние до которой также может служить мерой аномальности данных.

В данной статье рассмотрены варианты применения методов машинного обучения для обнаружения аномалий в навигационных данных, характеризующих параметры движения судна. Модельный подход используется для прогнозирования потребления топлива и поиска аномальных режимов.

Материалы и методы. Для верификации разработанных методов были использованы тестовые данные с судна, на котором в режиме тестовой эксплуатации установлена система сбора данных. На судне собирается информация с большого количество датчиков. В рамках данной работы используются значения только таких параметров, как скорости относительно воды (STW) и земли (SOG), курс (COG), угловая скорость (ROT), истинное значение курса (HDG), глубина (Depth), минимальная дистанция схождения с другим судном(CPA) и время до него (TCPA). Также для моделирования потребления топлива используются данные по его расходу.

В рамках тестирования алгоритмов использованы данные, собранные за месяц и представимые в виде множества

$$X = \left\{\left(t_i, x_{1,i}, x_{2,i}, \ldots, x_{9,i}\right)\right\}_{i=1}^{N}, \quad (1)$$

где t_i обозначает время поступления данных, а $x_{1,i}, x_{2,i}, \ldots, x_{9,i}$ – значения параметров, указанных выше в той же последовательности в момент времени t_i. Величина N определяет общее число накопленных данных. Заметим, что разные параметры измеряются датчиками с различной частотой. Для нормализации данных используется простая линейная интерполяция по времени

$$\frac{x_{k,j} - x_{k,i}}{x_{k,i+1} - x_{k,i}} = \frac{t_j - t_i}{t_{i+1} - t_i}, k = \overline{1,9},$$

такая, что соседние интервалы по времени равны постоянной величине $\Delta t = t_j - t_{j-1} =$ 1 мин. Таким образом, в каждый момент времени рассматривается девятимерный вектор состояния судна.

В качестве основного метода поиска аномалий использован алгоритм изолирующего леса. Модель прогнозирования потребления топлива построена на основе полиномиальной регрессии. Алгоритмы реализованы на языке Python с использованием библиотеки scikit-learn. На некоторых приведенных в статье рисунках на осях отсутствуют числовые значения, что обусловлено нормировкой реальных значений в диапазоне [0; 1].

Следует отметить, что для более эффективной работы алгоритмов поиска аномалий, а также для наглядной визуализации применен метод главных компонент, основная идея которого заключается в поиске ортогональных проекций, на которые разброс данных максимален [10]. В этом случае множество (1) преобразуется к

$$X = \left\{\left(t_i, PC_{1,i}, PC_{2,i}, \ldots, PC_{9,i}\right)\right\}_{i=1}^{N}, \quad (2)$$

где $PC_{k,i}$ обозначает k-ую главную компоненту вектора состояния X в момент времени i.

Результаты. Основную идею поиска опасных маневров можно продемонстрировать на примере анализа соотношений линейной и угловой скоростей судна. В нормальном режиме при значительных линейных скоростях судно должно двигаться прямолинейно. Случаи, когда у судна на высокой скорости наблюдается большая угловая скорость, соответствуют резким маневрам. На Рисунке 1 отображено подмножество множества точек (1) $\left\{\left(x_{1,i}, x_{4,i}\right)\right\}$.

В общем случае на маневры судна оказывают влияние как минимум все указанные выше параметры. Так, например, с точки зрения безопасности не желательно движение с большой скоростью на малых глубинах. Также на маневры оказывает влияние наличие рядом с судном других судов, которое может быть определено по величинам CPA и TCPA. На Рисунке 2 представлен результат работы алгоритма поиска аномалий на всем наборе входных данных, преобразованных методом главных компонент (2). Для визуализации использованы только три первые главные компоненты $\left\{\left(PC_{1,i}, PC_{2,i}, PC_{2,i}\right)\right\}$.

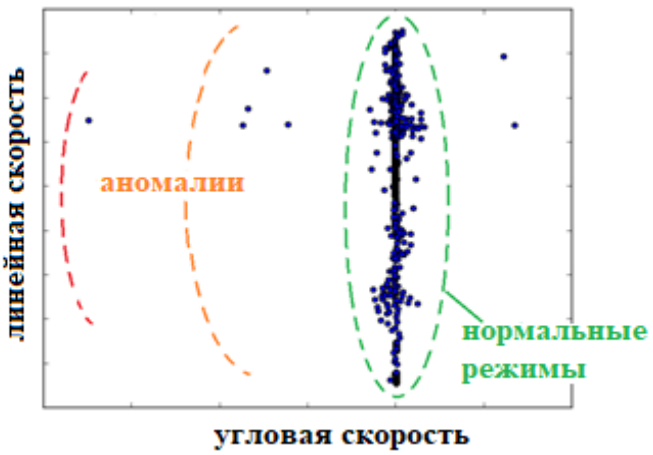

Рисунок 1 – Аномальные маневры в пространстве линейной и угловой скоростей

Отметим, что на Рисунке 2 можно увидеть несколько кластеров, описывающих маневры. Возможным объяснением таких групп может служить наличие различных режимов движения судна на открытой воде, в порту, при поворотах. Детальное рассмотрение и анализ обнаруженных кластеров выходит за рамки данной работы.

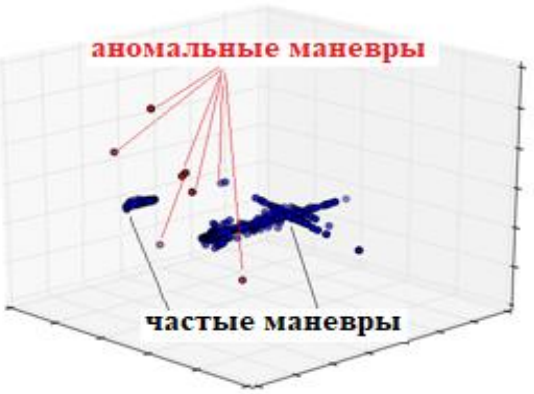

Рисунок 2 – Аномальные маневры в пространстве главных компонент

Найденные алгоритмом изолирующего леса аномалии могут быть ранжированы по близости к разделяющей гиперплоскости либо при помощи использования метрических методов. Данное значение можно интерпретировать как вероятность аномалии и использовать в качестве параметра для настройки количества обнаруженных аномалий за определенный промежуток времени. В данной работе использовалась встроенная в библиотеку scikit-learn функция $f(x) = decision\_function(x)$, вычисляющая для каждой точки множества (1) “меру нормальности” как глубину листа, содержащего эту точку, что эквивалентно количеству разветвлений, необходимых для ее изоляции:

$$F(X) = \left\{t_i,\ f(x_{1,i}, x_{2,i}, \ldots, x_{9,i})\right\}_{i=1}^{N}, \qquad (3)$$

Несмотря на то, что методы поиска аномалий изначально рассматриваются как полностью автоматические, предметная область может накладывать определенные ограничения. Так, например, при анализе маневров с точки зрения безопасности движения особый интерес представляют только относительно большие скорости судна. В данном случае практически невозможно добиться от метода автоматического извлечения подобной информации из данных. Требуется либо тонкая настройка параметров алгоритма, либо предварительная трансформация данных. Так, на Рисунке 3 представлен результат работы алгоритма изолирующего леса на множестве точек $\left\{\left(x_{1,i}, \frac{x_{4,i}}{x_{1,i}}\right)\right\}$. Здесь величина $\frac{x_{4,i}}{x_{1,i}}$ равна отношению угловой и линейной скоростей и определяет кривизну траектории. Движение с нулевой кривизной соответствует движению по прямой. Красным цветом обозначены точки, в которых значение нормированной в интервале [0; 1] функции f больше 0,95. Из рисунка видно, что аномалии обнаруживаются в области низких скоростей.

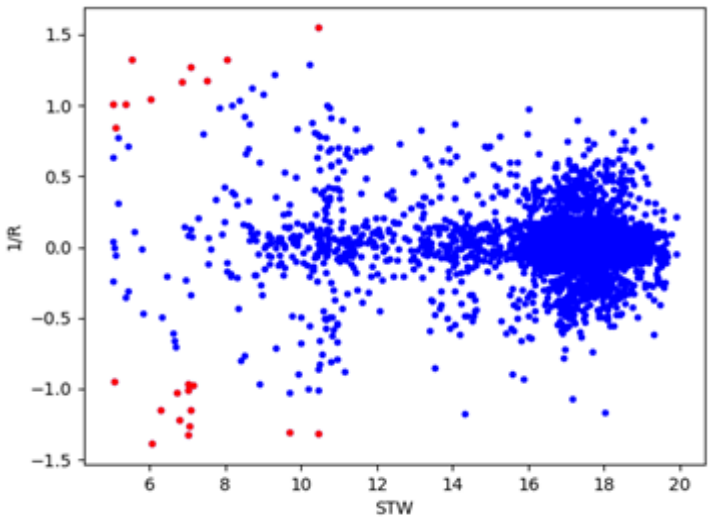

Рисунок 3 – Аномальные маневры в пространстве скорости (STW) и кривизны траектории (R – мгновенный радиус поворота)

Для решения таких проблем требуется предварительная фильтрация данных. Например, на Рисунке 4 приведена работа алгоритма изолирующего леса на данных, первоначально подвергнутых экспертной фильтрации

$$\tilde{X} = \left\{ \left(t_i, x_{1,i}, x_{2,i}, \ldots, x_{9,i} \right) \middle| x_{1,i} > 12 \right\}_{i=1}^{N}.$$

Зеленым цветом обозначены точки, которые не интересны в задаче обнаружения аномалий ввиду невысокой скорости движения судна. Синим – нормальные ситуации маневров. Красным и оранжевым отмечены подозреваемые в аномальности объекты с высоким значением функции f.

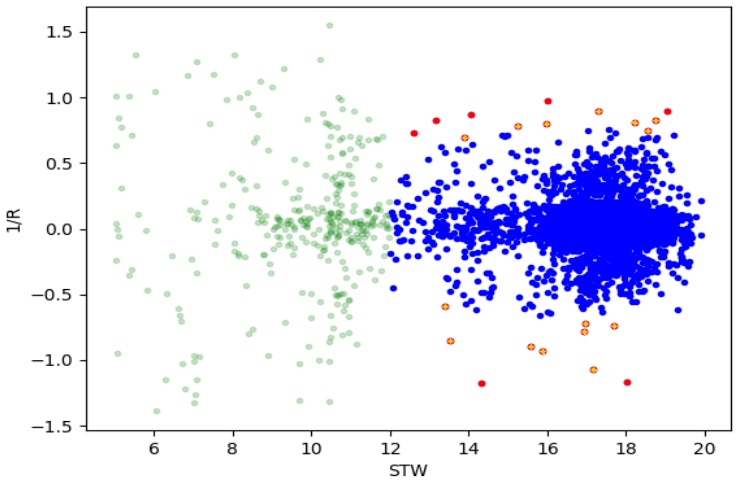

Рисунок 4 – Аномальные маневры в пространстве скорости (STW) и кривизны траектории (R – мгновенный радиус поворота) с предварительной фильтрацией

Найденные аномалии могут быть отображены и на маршруте судна в географических координатах. На Рисунке 5 представлен маршрут с обнаруженными выше аномалиями маневров. Отметим, что резкие маневры могут быть вызваны разными причинами, такими как уход от столкновения, корректировка маршрута и др. Кроме того, возможны и выбросы в данных, когда аномалии являются случайным шумом. Предполагается, что анализ причин аномальных ситуаций в дальнейшем будет производиться дополнительно с использованием статистических тестов и экспертной оценки.

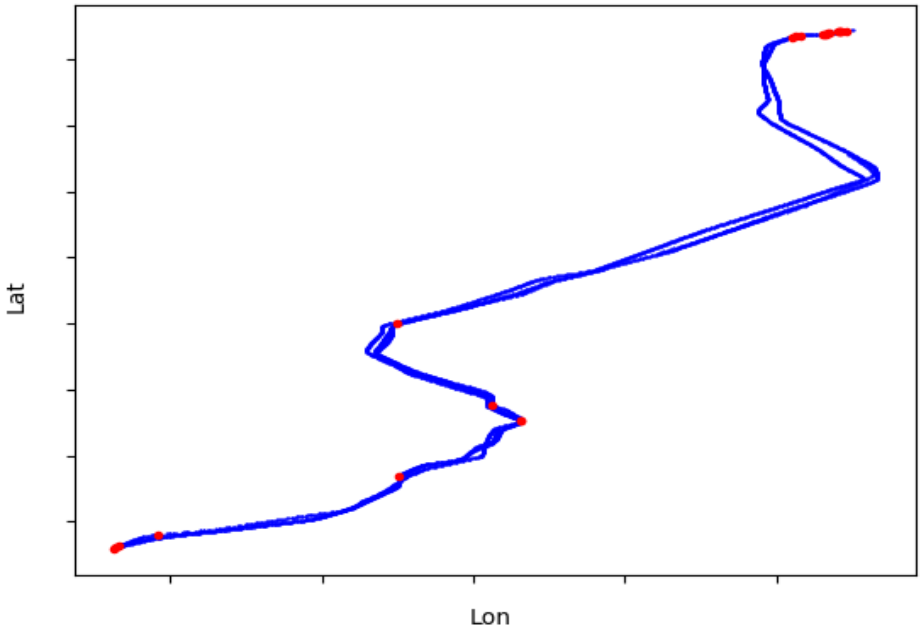

Рисунок 5 – Аномальные маневры в пространстве скорости (STW) и кривизны траектории (R – мгновенный радиус поворота) с привязкой к географическим координатам (Lon – долгота, Lat – широта)

Рассмотренный подход для обнаружения аномалий на основе машинного обучения без учителя может быть легко обобщен на другие случаи, связанные с опасными параметрами качки судна, неправильными маневрами или опасной глубиной. Для обнаружения аномалий потребления топлива предлагается подход обучения с учителем, когда в данных множества (1) восстанавливается зависимость расхода топлива по навигационным данным судна $\{x_{1,i}, x_{2,i}, \ldots, x_{8,i}\} \to x_{9,i}$. В качестве модели используется полиномиальная регрессии, которая в общем случае может быть записана в виде

$$y_i = A_0 + (x_{1,i}, x_{2,i}, \ldots, x_{8,i})A_1 + (x_{1,i}{}^2, x_{1,i}x_{2,i}, \ldots, x_{2,i}{}^2, x_{2,i}x_{3,i}, \ldots, x_{8,i}{}^2)A_2 + \ldots, \quad (4)$$

до требуемого порядка нелинейности. Здесь A_k обозначает вектор-столбец числовых параметров с размерностью, соответсвующей вектору нелинейностей порядка k. Вычисление этих параметров осуществляется

методом наименьших квадратов, минимизирующий среднеквадратичную ошибку отклонения прогноза y_i от реального расхода топлива на милю $x_{9,i} / x_{1,i}$.

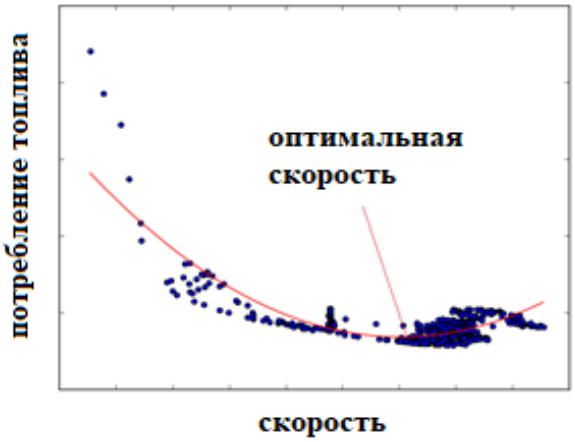

Рисунок 6 – Модель потребления топлива на основе квадратичной зависимости от скорости движения судна

Использование полиномиальной регрессии обусловлено хорошей интерпретируемостью модели. Так, уже даже в случае применения простой квадратичной функции ($k=2$ в формуле (4)) удается найти, например, оптимальную скорость движения судна по отношению к потребляемому объему топлива. На рис. 6 представлен график зависимости расхода топлива на милю от скорости движения судна.

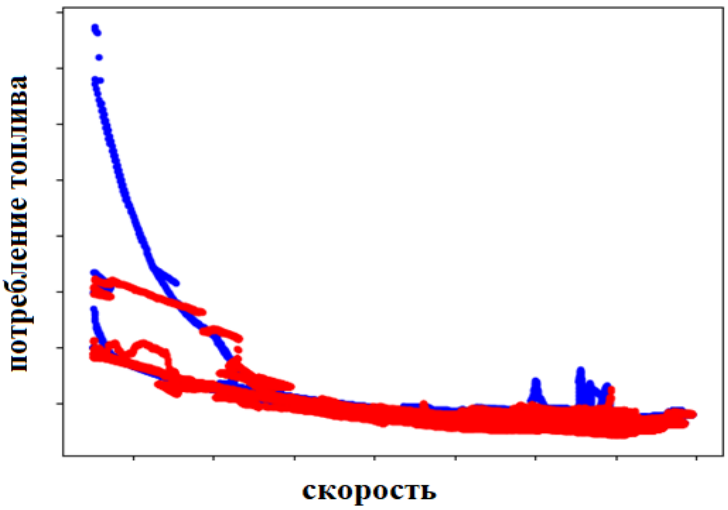

Рисунок 7 – Аномальные маневры в пространстве скорости (STW) и кривизны траектории (R – мгновенный радиус поворота)

Для повышения точности модели была применена многофакторная полиномиальная регрессия третьего порядка. Данная модель учитывает не только скорость движения судна, но и глубину, погодные данные (ветер, течение), угол дрейфа судна. Данные характеристики вычисляются на основе параметров $(x_{1,i}, x_{2,i}, \ldots, x_{8,i})$, а затем нормируются в интервал [0, 1].

Для построения модели использовалось разбиение на обучающую и тестовую выборки в соотношениях 70% и 30% соответственно. Средняя ошибка (квадратный корень из среднеквадратической ошибки) на тестовых данных составила 6%, причем 95% точек пространства имеют ошибку, меньшую 13,5%. На рис. 7 синим цветом обозначены реальные данные по потреблению топлива, красным – значения, спрогнозированные моделью. Построенную модель потребления топлива можно применять и для вычисления интегральной величины потребления топлива за маршрут. Для этого прогнозные значения расхода топлива на милю y_i должны быть проинтегрированы вдоль всего маршрута движения судна. В дискретном виде формула для вычисления расхода топлива за маршрут запишется в виде

$$Y = \sum_{i=1}^{T} y_i x_{1,i} (t_i - t_{i-1}) \tag{5}$$

При использовании формулы (5) для оценки точности модели потребления топлива получено, что в 95% случаев ошибка на маршруте не превышает 6%.

По Рисунку 7 можно заметить, что есть случаи, которые плохо объясняются моделью. Такое аномальное поведение требует дальнейшего рассмотрения и выяснения причин возникновения расхождения. В качестве теоретических причин можно назвать несанкционированный отбор топлива либо нештатный режим работы механизмов, приводящий к повышенному потреблению.

Обсуждение. Рассмотренные подходы к обнаружению аномалий обладают достаточной степенью универсальности и могут быть применены для различных сценариев. Однако, как продемонстрировано в данной работе, перед исследованием все равно возникает необходимость предобработки данных и, как минимум, указания областей в пространстве данных, в которых следует ожидать аномальное поведение.

Основными направлениями дальнейшего развития подходов являются разработка методов объяснения найденных аномалий и их фильтрация от случайных выбросов. Здесь можно выделить два подхода. К первому относятся такие статистические методы, как факторный анализ и проверка гипотез. Второй метод основан на машинном обучении с учителем и требует сбора обратной связи по найденным аномалиям с

классификацией типов каждого случая. Этот вариант более трудоемок, так как требует сбора маркированных данных.

Для повышения точности модели прогнозирования потребления топлива можно предложить методы, основанные на градиентном бустинге моделей. Другим подходом является построение физической модели движения судна, а также учет параметров работы двигателя.

Заключение. В рамках исследования продемонстрировано применение метода изолирующего леса для задачи обнаружения аномалий маневрирования морских судов. Стандартная реализация алгоритма взята из библиотеки scikit-learn и модифицирована с учетом требований предварительной фильтрации входных данных. Построена и протестирована модель прогнозирования потребления топлива на основе навигационных данных.

В работе определен ряд трудностей, возникающих в процессе применения алгоритмов, приведены варианты их решения и возможные пути дальнейшего развития подходов. Реализация этих методов позволит в будущем перейти к задачам как анализа больших данных, собираемых с судов, так и к внедрению подобных алгоритмов непосредственно на борту судна с целью уменьшения передаваемого объема информации.

Автор выражает благодарность Иванову А.Н. за помощь в получении тестовых данных, а также проф. Андрианову С.Н. за консультации в вопросах применения рассмотренных методов.

ЛИТЕРАТУРА

1. А.С. Пинский. Е-Навигация и безэкипажное судовождение // Транспорт РФ. Журнал о науке, практике, экономике. 2016. №4 (65). С. 50–54.
2. Brandsæter, G, Manno , E. Vanem, I. Glad. An application of sensor-based anomaly detection in the maritime industry // IEEE International Conference on Prognostics and Health Management, 2016. P. 1–8.
3. Y. Liu,W. Ding. A KNNS based anomaly detection method applied for UAV flight data stream // *Prognostics and System Health Management Conference*, Beijing, 2015. doi: 10.1109/PHM.2015.7380051.
4. K. D. Borne. Effective Outlier Detection using K-Nearest Neighbor Data Distributions: Unsupervised Exploratory Mining of Non-Stationarity in Data Streams. URL: https://pdfs.semanticscholar.org/f3eb/4573d3164345063351979c9409014ec33d4d.pdf
5. В.П. Шкодырев, К.И. Ягафаров, В.А. Баштовенко, Е.Э. Ильина. Обзор методов обнаружения аномалий в потоках данных // Proc. of

the Second Conference on Software Engineering and Information Management, Санкт-Петербург, Россия, 2017. Vol. 1864.

6. Д.В. Заварзин. К вопросу поиска аномалий во временных рядах // Инновации в науке: сб. ст. по матер. XXIX междунар. науч.-практ. конф. № 1(26). – Новосибирск: СибАК, 2014. С. 59–64.
7. T. Klerx, M. Anderka, H. K. Büning, S. Priesterjahn, Model-Based Anomaly Detection for Discrete Event Systems // *IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, 2014, pp. 665-672. doi: 10.1109/ICTAI.2014.105
8. L. Simon, A.W. Rinehart. A Model-Based Anomaly Detection Approach for Analyzing Streaming Aircraft Engine Measurement Data. URL: ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150000721.pdf.
9. T. Liu, K. M. Ting, Zhi-Hua Zhou. Isolation Forest. URL: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf.
10. F. Harrou, F. Kadri, S. Chaabane, C. Tahon, Y. Sun. Improved principal component analysis for anomaly detection: Application to an emergency department // Computers & Industrial Engineering, 88, 2015. P. 63–77.

A.A. Sholokhova

ANOMALY DETECTION IN SENSOR DATA IN APPLICATION TO THE ANALYSIS OF MARITIME VESSEL MOTION

Saint Petersburg State University

The article describes algorithms for anomalies detection in the sensory data in the application to the analysis of maritime traffic. The modern vessel is equipped with many sensors, continuously recording the performance of its various subsystems. The collection and storage of such information provide the possibilities of using intelligent data analysis systems. These tasks concern both the issues of ensuring maritime traffic safety (analysis and prevention of dangerous maneuvers) and the problems of increasing economic efficiency (increased fuel consumption) for ship-owners. The onboard sensors can generate numerical data with a frequency from ten seconds for telemetry up to one minute for navigational parameters. Given such a large amount of information, it becomes obvious the need for the development of automated decision support systems. Applied areas of such systems can serve tasks of preventing dangerous maneuvers, predicting maintenance, preventing collisions, optimizing fuel consumption. In the article, the application unsupervised learning for the analysis of navigational data (ship coordinates, speed, course, depth, etc.) and an example of predicting of fuel consumption based on regression models are considered. A description of various mathematical approaches and its demonstration on real data is given. In conclusion, the possible development and improvement of the given methods are considered.

Keywords: anomaly detection, sensor data, extreme maneuvering, prediction of fuel consumption

REFERENCES

1. A. Pinskiy. E-Navigation and unmanned ship navigation. Transport of RF. Journal on Science, Practice, and Economics. 2016. N4 (65). P. 50-54.
2. Brandsæter, G, Manno , E. Vanem, I. Glad. An application of sensor-based anomaly detection in the maritime industry // IEEE International Conference on Prognostics and Health Management, 2016. P. 1–8.
3. Y. Liu,W. Ding. A KNNS based anomaly detection method applied for UAV flight data stream // *Prognostics and System Health Management Conference*, Beijing, 2015. doi: 10.1109/PHM.2015.7380051.
4. K. D. Borne. Effective Outlier Detection using K-Nearest Neighbor Data Distributions: Unsupervised Exploratory Mining of Non-Stationarity in Data Streams. URL: https://pdfs.semanticscholar.org/f3eb/4573d3164345063351979c9409014ec33d4d.pdf
5. V. Shkodyrev, K. Yagafarov, V. Bashtovenko, E. Ilyina. The Overview Of Anomaly Detection Methods in Data Streams // Proceedings of the Second Conference on Software Engineering and Information Management, Saint Petersburg, Russia, April 21, 2017. Vol. 1864.
6. D. Zavarzin. About anomalies detection techniques in time series // Innovations in science: Proc. on XXIX Intern. Conf № 1(26). – Novosibirs, 2014. P. 59–64.

7. T. Klerx, M. Anderka, H. K. Büning, S. Priesterjahn, Model-Based Anomaly Detection for Discrete Event Systems // *IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, 2014, pp. 665-672. doi: 10.1109/ICTAI.2014.105
8. L. Simon, A.W. Rinehart. A Model-Based Anomaly Detection Approach for Analyzing Streaming Aircraft Engine Measurement Data. URL: ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150000721.pdf.
9. T. Liu, K. M. Ting, Zhi-Hua Zhou. Isolation Forest. URL: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf.
10. F. Harrou, F. Kadri, S. Chaabane, C. Tahon, Y. Sun. Improved principal component analysis for anomaly detection: Application to an emergency department // Computers & Industrial Engineering, 88, 2015. P. 63–77.