Повышение точности поиска аномалий в данных на основе ансамбля моделей

Ю.А. Ханова, Н.В. Шевская

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

Аннотация: В статье рассмотрен процесс разработки и конфигурирования ансамбля моделей для поиска аномалий в данных. Предложена структурная схема ансамбля моделей и блок-схема алгоритма поиска аномалий в данных. Показано повышение точности поиска аномалий при использовании ансамбля моделей.
Ключевые слова: аномалия, алгоритм, ансамбль моделей, точность, матрица путаницы, набор данных, комитет, кластеризация, консольное приложение, модель.

Сегодня поиск аномалий актуален в самых разнообразных отраслях: в медицине – для выявления злокачественных опухолей при наличии аномалий на МРТ, в киберфизических системах – при выявлении дефектов в работе оборудования, в банковской сфере аномалии в данных транзакций по кредитной карте могут указывать на кражу личных данных, а в розничной торговле – при работе с клиентской базой, где такие параметры, как возраст, количество заказов или время, проведенное на сайте, могут находиться в определенных пределах [1-3]. В нескольких исследовательских сообществах были разработаны методы обнаружения аномалий, применяемые, как правило, отдельно друг от друга [4].

По прогнозам IDC глобальная информационная сфера данных к 2025 году вырастет более чем на порядок по сравнению с 2015 годом, с 17 до 175 ЗБ [5]. С увеличением объемов данных в процессе решения сложных задач классификации точность выявления аномалий может снижаться, а большинство известных методов становятся неактуальными. Обычно базовые модели (слабые ученики) работают отдельно хуже потому, что они имеют:

– высокое *смещение* ε_b (систематическая ошибка конкретного обучающего алгоритма). Смещение есть мера стабильности ошибки данного

алгоритма обучения, которая не может быть исключена даже применением бесконечного числа обучающих множеств. На практике такая ошибка не может быть вычислена точно, а оценивается только приблизительно.

– высокую *дисперсию* метода обучения для этой задачи ε_D. Обучающее множество используется частично — оно всегда конечно — и, следовательно, не полностью представляет реальную популяцию наблюдений.

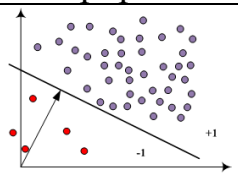
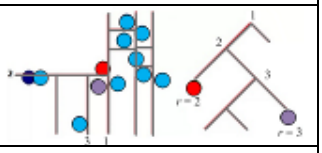
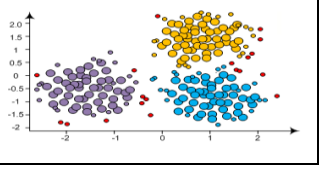
Общая ожидаемая ошибка базовой модели ε состоит из суммы дисперсии и смещения: $\varepsilon = \varepsilon_b + \varepsilon_D$. Композиция нескольких базовых моделей позволяет уменьшить ошибку за счет дисперсии. При этом, чем больше базовых моделей используется, тем меньше дисперсия. В этой связи получила распространение технология создания сильного ученика или композиций (ансамблей) из базовых моделей поиска аномалий, что обеспечивает взаимную компенсацию ошибок каждого из входящих в ансамбль алгоритмов [6]. При композиции обучающих алгоритмов используют различные методы поиска аномалий, варьирования структуры базовых моделей, способы агрегации результатов, что обеспечивает огромное разнообразие вариантов ансамблей моделей. Проанализированы методы определения аномалий по режиму распознавания: с учителем, без учителя, частично с учителем; по способу вычисления на основе: статистического анализа, нечеткой логики, машинного обучения, индуктивного вывода, генетических алгоритмов, искусственных нейронных сетей и гибридные методы [2].

В результате были выбраны три метода для поиска аномалий (таблица №1). Метод One Class SVM был выбран благодаря функции ядра – могут проводиться нелинейные границы. В отличии от других методов, для обучения One Class SVM не требуется наличие большого количества аномалий в обучающей выборке, что обычно необходимо для стандартной бинарной классификации [7]. Основным преимуществом метода Isolation Forest является возможность использования методов выборки в такой

степени, которая не допускается профильными методами, создавая очень быстрый алгоритм с низкой потребностью памяти [8]. Метод DBSCAN показывает более точные результаты, так как наборы данных неоднородны и определение областей плотности позволяет качественно отделить кластеры нормальных значений от аномалий, если параметры работы алгоритма подобраны правильно [9]. Далее необходимо выбрать способы обобщения результатов этих отдельных методов выявления аномалий.

Таблица №1

Методы поиска аномалий

Метод	Идея алгоритма	Интерпретация
Опорных векторов для одного класса One Class SVM	Разделяет классы гиперплоскостью с целью сделать расстояние между ними максимальным	
Изолирующего леса Isolation Forest	Работает по принципу изоляции аномалий, окончательный результат усредняется	
Пространственной кластеризации DBSCAN	Осуществляет поиск областей с более высокой плотностью, разделенных областями с низкой плотностью	

Наиболее известны методы композиции в ансамбле: однородные ансамбли, «ручные методы», комитеты (голосования)/усреднение, включая бэггинг, кодировки/перекодировки ответов, бустинг, стекинг [6]. В результате сравнения методов ансамблирования был выбран метод комитетов, так как он позволяет совместить несколько разных независимых методов поиска аномалий, что обеспечивает повышение точности результатов определения аномалий и расширяет сферу применения разрабатываемого инструмента. Важный вопрос, который необходимо решить при построении ансамбля моделей на основе комитетов, касается метода комбинирования результатов, выданных отдельными моделями, выбор идет между следующими подходами: голосование, взвешенное

голосование, усреднение, взвешенное усреднение. Выбранные методы поиска аномалий (таблица №1) объединены в ансамбль моделей, где общий результат определяется голосованием среди результатов отдельных моделей – слабых учеников. Решение для каждого экземпляра тестовой выборки принимается на основе выбора большинства, то есть, если две или более моделей определили экземпляр, как аномалию, то этот экземпляр является аномалией для результата ансамбля моделей. Структурная схема ансамбля моделей для поиска аномалий в данных на основе методов One Class SVM, Isolation Forest, DBSCAN представлена на рис. 1.

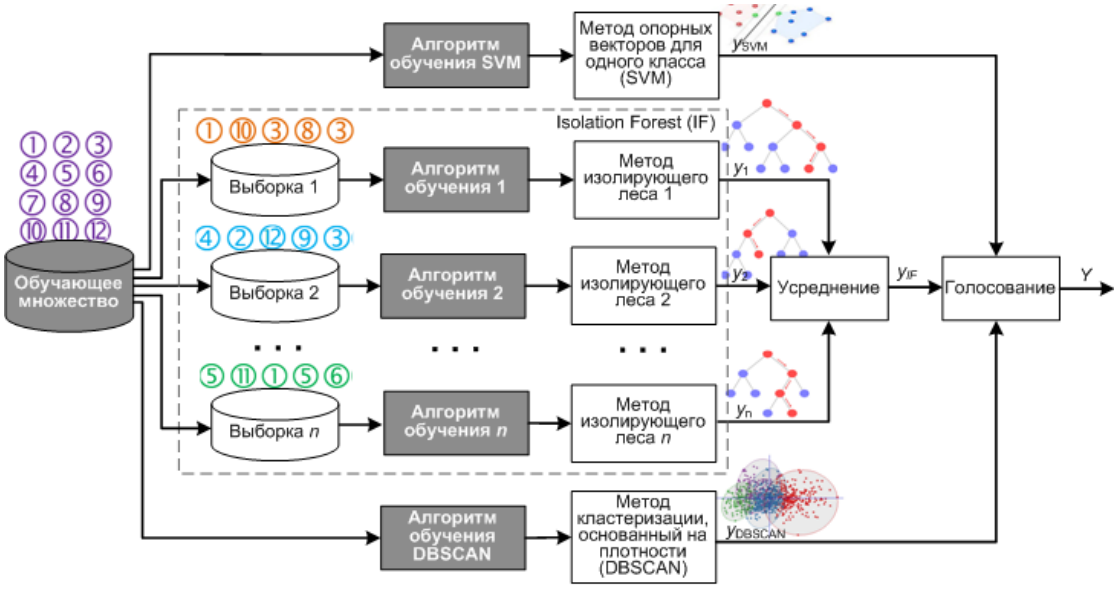


Рис. 1. – Структурная схема ансамбля моделей

Задачу поиска аномалий на основе ансамбля выбранных моделей можно представить в виде нескольких этапов: предварительная обработка данных, последовательная реализация каждого, из выбранных методов поиска аномалий (таблица №1), реализация ансамбля полученных моделей, определение метрик, необходимых для анализа результатов (рис. 2).

Консольное приложение для анализа точности результатов поиска аномалий отдельными методами и ансамблем моделей разработано средствами языка программирования общего назначения Python 3, а также библиотеки машинного обучения Scikit-learn. Для обработки данных

используется библиотека Pandas, позволяющая работать с набором данных, как с многомерным массивом.

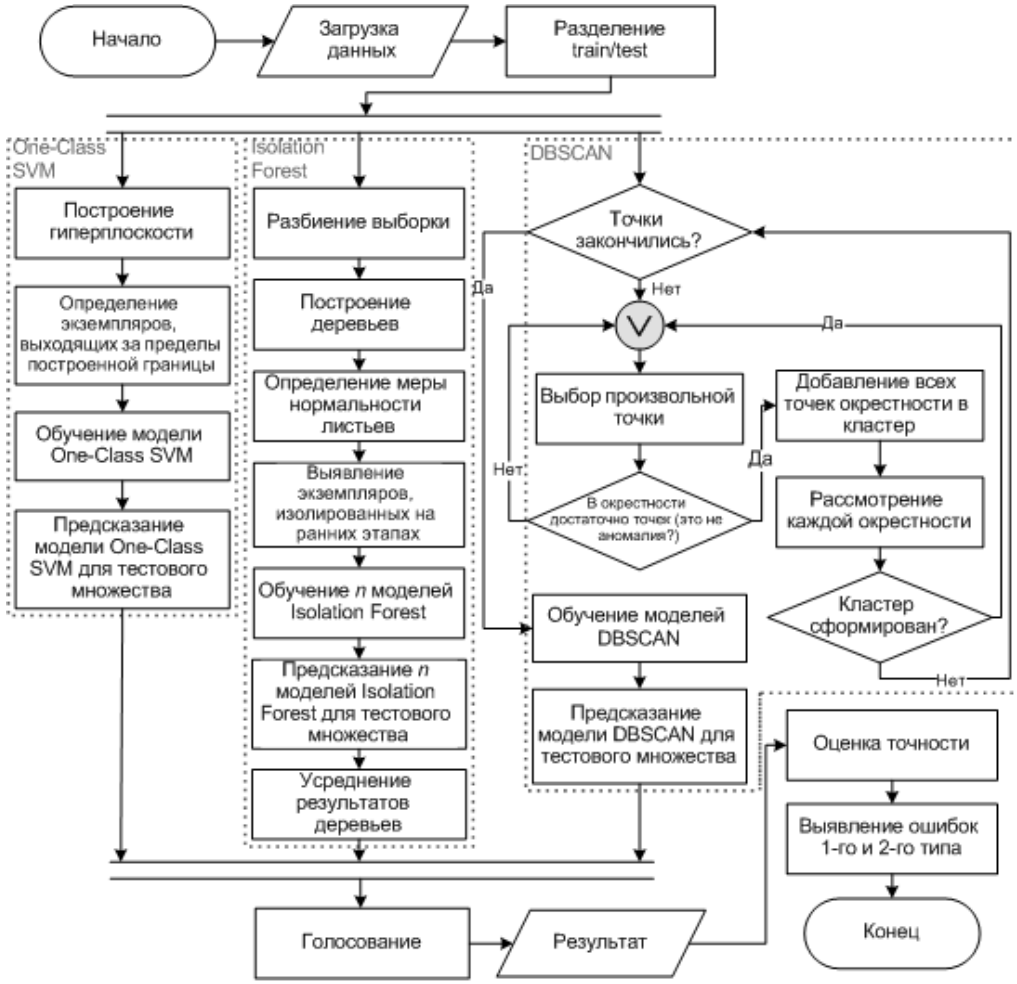


Рис. 2. – Алгоритм поиска аномалий

Для тестирования алгоритма (рис. 2) возьмем набор данных Supermarket sales, содержащий исторические данные о продажах компании-супермаркета, которые были зарегистрированы в 3 разных филиалах за 3 месяца (рис. 3). После чтения файла происходит разделение исходных данных на тренировочную и тестовую выборки [10]. В качестве признаков 1-4 взяты поля 7-10 набора данных соответственно.

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	01.05.2019
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	03.08.2019
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	03.03.2019
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	27.01.2019

Рис. 3. – Структура набора данных Supermarket sales (фрагмент данных)

В качестве метрик для оценивания результатов работы алгоритмов были выбраны точность, количество выявленных аномалий, количество ошибок первого и второго типа. Для получения этих значений была использована матрица путаницы, где сопоставляются фактические и полученные данные. Матрица путаницы позволяет измерять производительности для классификации машинного обучения. Вычисление матрицы путаницы дает лучшее представление о том, что классификационная модель делает правильно и какие типы ошибок она допускает. В данной задаче матрица путаницы имеет размер 2×2 и включает 4 составляющие: *TP* (True Positive) - определено как аномалия и фактически является аномалией; *FN* (False Negative) - определено как аномалия, но фактически является нормальным значением (ошибка 1-го типа); *TN* (True Negative) - определено как нормальное значение и фактически является нормальным значением; *FP* (False Positive) - определено как нормальное значение, но фактически является нормальным значением (ошибка 1-го типа). Точность ε вычислена с помощью матрицы путаницы по формуле:

$$\varepsilon = \frac{TP + TN}{TP + FP + TN + FN}.$$

Каждая модель в ансамбле отдельно обучена, после обучения модели проведено предсказание на тестовых данных. На выходе каждому экземпляру выборке присваивается значения: «1» – нормальное значение и «-1» – аномальное значение (рис. 4). В ансамбле моделей (Ensemble) решение принимается на основе выбора большинства, если две или более моделей определили экземпляр, как аномалию, то этот экземпляр является аномалией для ансамбля.

```
Ensamble:
[ 1 -1  1  1  1  1  1 -1  1  1  1  1 -1  1 -1  1 -1  1  1  1  1  1 -1
  1 -1  1  1  1 -1  1 -1  1  1  1  1  1  1  1  1  1  1  1 -1  1  1  1
 -1  1  1]
error type 1:  2
error type 2:  4
number of detected anomalies:  11
accuracy:  88.24%
```

Рис. 4. – Результат выполнения программы для ансамбля моделей

Получены графики, отображающие набор данных и аномалии, которые удалось определить (рис. 5), матрица путаницы (рис. 6), а также графики значений для каждого отдельного признака с спроецированными на них аномалиями (рис. 7). Графики и матрицы путаницы (рис. 5-7) построены для каждой модели отдельно и для ансамбля в целом.

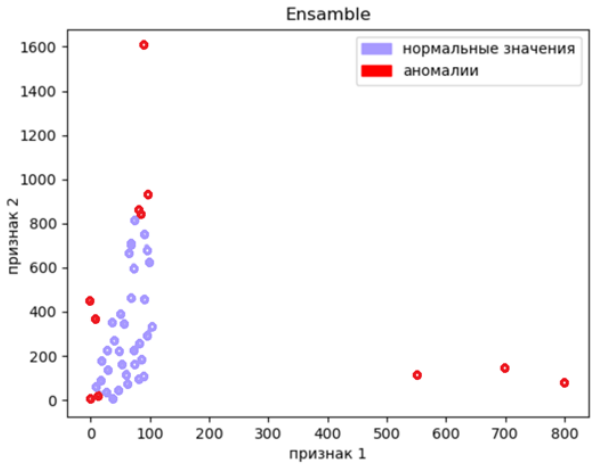


Рис. 5. – Аномалии, определенные ансамблем

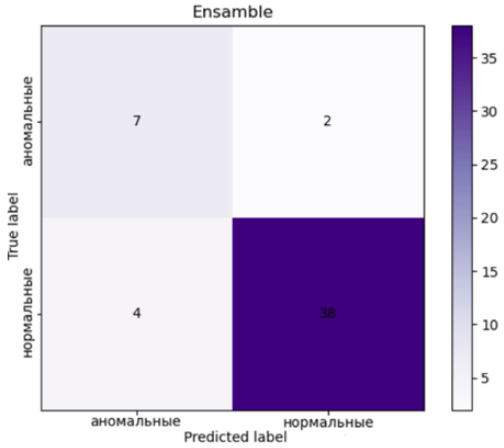


Рис. 6. – Матрица путаницы ансамбля

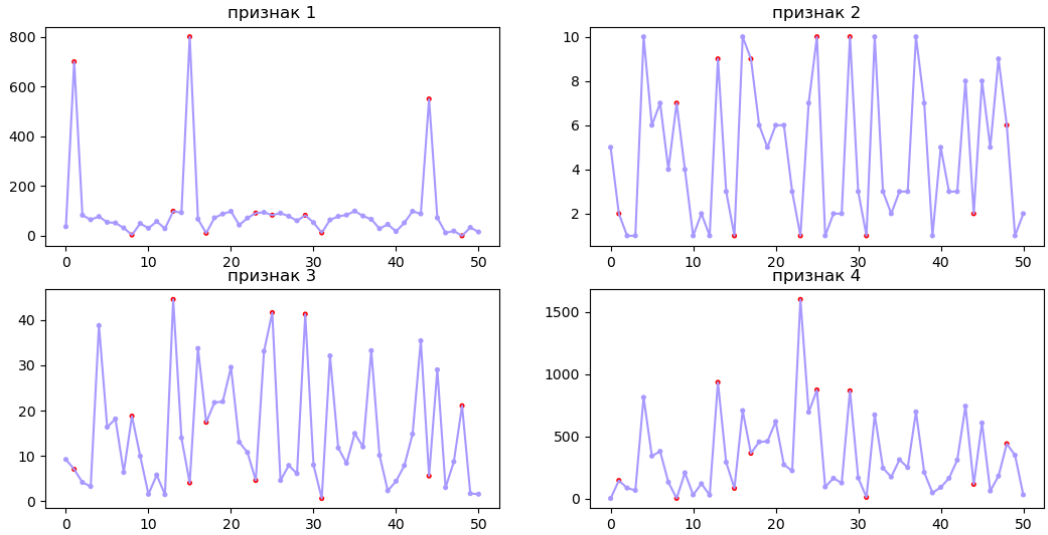


Рис. 7. – Графики значений отдельных признаков ансамбля моделей

Сравнение результатов работы отдельных методов и ансамбля представлены в таблице №2. Фактическое количество аномалий в тестовом наборе – 9. Как видно из таблицы №2, точность результатов ансамбля моделей выше, чем точность результатов отдельных моделей.

Таблица № 2

Сравнение результатов работы программы

Метрика / Метод	Количество выявленных аномалий	Количество ошибок 1-го типа	Количество ошибок 2-го типа	Точность определения аномалий
One Class SVM	15	1	7	84.31%
Isolation Forest	12	3	6	82.35%
DBSCAN	7	5	3	84.31%
Ensemble	11	2	4	88.24%

Детально рассмотрена процедура создания, конфигурирования и тестирования ансамбля моделей для поиска аномалий в данных. Значение точности, полученное в результате тестирования разработанного ансамбля моделей для поиска аномалий, оказалось выше, чем точность отдельных моделей, и составило 88,24%.

Литература

1. Ажмухамедов И.М., Марьенков А.Н. Поиск и оценка аномалий сетевого трафика на основе циклического анализа // Инженерный вестник Дона, 2012, №2. URL: ivdon.ru/ru/magazine/archive/n2y2012/742.

2. Chandola, V., Banerjee, A., Kumar V. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 p.

3. Protalinskiy O., Savchenko N., Khanova A. Data mining integration of power grid companies enterprise asset management. Studies in Systems, Decision and Control. 2020. T. 260. pp. 39-49.

4. Hodge, V., Austin, J. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review. 2004. 22. pp. 85-126.

5. Reinsel D., Gantz J., Rydning J. Data Age 2025: The Evolution of Data to Life-Critical. Don't Focus on Big Data; Focus on the Data That's Big // An IDC White Paper. 2017. 25 p.

6. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям: Учебное пособие. СПб.: Питер, 2010. 704 с.

7. Зубков Е. В., Белов В. М. Методы интеллектуального анализа данных и обнаружение вторжений // Вестник СибГУТИ. 2016. № 1. С. 118-133.

8. Liu F. T., Ting K. M., Zhou Z. H. Isolation Forest. Data Mining, ICDM'08. Eighth IEEE International Conference on. IEEE, 2008. pp. 413-422.

9. Чесноков М.Ю. Поиск аномалий во временных рядах на основе ансамблей алгоритмов DBSCAN // Искусственный интеллект и принятие решений. 2018. № 1. С. 99-107.

10. Лила В.Б. Алгоритм и программная реализация адаптивного метода обучения искусственных нейронных сетей // Инженерный вестник Дона, 2012, №1. URL: ivdon.ru/ru/magazine/archive/n1y2012/626.

References

1. Azhmukhamedov I.M., Maryenkov A.N. Inzhenernyj vestnik Dona, 2012, №2. URL: ivdon.ru/ru/magazine/archive/n2y2012/742.

2. Chandola, V., Banerjee, A., Kumar V. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 p.

3. Protalinskiy O., Savchenko N., Khanova A. Studies in Systems, Decision and Control. 2020. T. 260. pp. 39-49.

4. Hodge, V., Austin, J. Artificial Intelligence Review. 2004. 22. pp. 85-126.

5. Reinsel D., Gantz J., Rydning J. An IDC White Paper. 2017. 25 p.

6. Paklin N.B., Oreshkov V.I. Biznes-analitika: ot dannyh k znanijam. [Business analytics: from data to knowledge]. SPb.: Piter, 2010. 704 p.

7. Zubkov E.V., Belov V.M. Vestnik SibGUTI. 2016. № 1. pp. 118-133.

8. Liu F. T., Ting K. M., Zhou Z. H. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008. pp. 413-422.

9. Chesnokov M.Yu. Iskusstvennyj intellekt i prinyatie reshenij.. 2018. № 1. pp. 99-107.

10. Lila V.B. Inzhenernyj vestnik Dona, 2012, №1. URL: ivdon.ru/ru/magazine/archive/n1y2012/626.