МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ВСЕРОССИЙСКАЯ ГРУППА ТЕОРИИ ИНФОРМАЦИИ IEEE
АКАДЕМИЯ ИНФОРМАТИЗАЦИИ ОБРАЗОВАНИЯ
ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ООО «ОТКРЫТЫЕ РЕШЕНИЯ»
ОБЩЕСТВО «ЗНАНИЕ» РОССИИ
ПРИВОЛЖСКИЙ ДОМ ЗНАНИЙ

XXII Международная
научно-техническая конференция

ПРОБЛЕМЫ ИНФОРМАТИКИ В ОБРАЗОВАНИИ, УПРАВЛЕНИИ, ЭКОНОМИКЕ И ТЕХНИКЕ

Сборник статей

Декабрь 2022 г.

Пенза

УДК 004
ББК 32.81я43+74.263.2+65.050.2я43
П781

П781 **ПРОБЛЕМЫ ИНФОРМАТИКИ В ОБРАЗОВАНИИ, УПРАВЛЕНИИ, ЭКОНОМИКЕ И ТЕХНИКЕ :**
сборник статей XXII Международной научно-технической конференции. – Пенза: Приволжский Дом знаний, 2022. – 356 с.

ISBN 978-5-8356-1800-2
ISSN 2311-0406

Под редакцией ***В.И. Горбаченко,*** доктора технических наук, профессора;
В.В. Дрождина, кандидата технических наук, профессора

Информация об опубликованных статьях предоставлена в систему Российского индекса научного цитирования (РИНЦ) по договору № 573-03/2014К от 18.03.2014.

ISBN 978-5-8356-1800-2
ISSN 2311-0406

XXII International
scientific and technical conference

PROBLEMS OF INFORMATICS IN EDUCATION, MANAGEMENT, ECONOMICS AND TECHNICS

December, 2022

Penza

Кревский М.И.
ГКУ «Информационный город»,
г. Москва, Россия

Krevskiy M.I.
State Government Institution
"Info City", Russia, Moscow

УДК 519.23

АНАЛИЗ АНСАМБЛЕВЫХ МЕТОДОВ КЛАССИФИКАЦИИ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОСЛЕОПЕРАЦИОННЫХ ОСЛОЖНЕНИЙ У БОЛЬНЫХ ЖЕЛЧНО-КАМЕННОЙ БОЛЕЗНЬЮ

Р. Н. Кузнецов, О. Ю. Кузнецова

ANALYSIS OF ENSEMBLE CLASSIFICATION METHODS TO PREDICT POSTOPERATIVE COMPLICATIONS IN PATIENTS WITH CHOLELITHIASIS

R. N. Kuznetsov, O. Yu. Kuznetsova

Аннотация. В этой работе описаны ансамблевые алгоритмы. Как правило, ансамблевые модели объединяют несколько базовых моделей для повышения производительности прогнозирования. Наиболее известным примером модели ансамбля является случайный лес, который, значительно упрощая логику алгоритма, объединяет несколько деревьев решений и агрегирует их прогнозы, используя большинство голосов.

Ключевые слова: прогнозирование послеоперационных осложнений, случайный лес, ROC-кривая, ансамбль голосования, деревья решений.

Abstract. In this paper, ensemble algorithms are described. As a rule, ensemble models combine several basic models to improve forecasting performance. The most famous example of an ensemble model is a Random Forest, which, greatly simplifying the logic of the algorithm, combines several decision trees and aggregates their predictions using a majority of votes.

Key words: prediction of postoperative complications, random forest, ROC curve, voting ensemble, decision trees.

Аналогично случайному лесу, ансамбль голосования оценивает несколько базовых моделей и использует голосование для объединения отдельных прогнозов, чтобы получить окончательный результат. Однако ключевое различие заключается в базовых оценках. Такие модели, как ансамбль голосования, не требуют, чтобы базовые модели были однородными. Другими словами, мы можем обучать разных базовых учащихся, например, дереву решений и логистической регрессии, а затем использовать ансамбль голосования для объединения результатов. Данный метод был использован для повышения точности прогнозирования путем объединения прогнозов нескольких моделей машинного обучения. Традиционный подход заключался в объединении так называемых «слабых» учащихся. Однако в данной работе реализован подход, заключающийся в создании ансамбля из хорошо подобранной коллекции сильных, но разнообразных моделей.

Классификатор голосования – это оценщик, который объединяет модели, представляющие различные алгоритмы классификации, связанные с индивидуальными весами для обеспечения достоверности. Оценщик классификатора голосования, построенный путем объединения различных моделей классификации, оказывается более сильным метаклассификатором, который уравновешивает слабые стороны отдельных классификаторов в конкретном наборе данных. Классификатор голосования принимает большинство голосов на основе весов, применяемых к классу или вероятностям классов, и присваивает записи метку класса на основе большинства голосов. Прогноз ансамбля классификаторов может быть математически представлен следующим образом:

$$\hat{y} = arg \max_{i} \sum_{j=1}^{m} w_j \, \chi_A\big(C_j(x) = i\big),$$

где *Cj* представляет классификатор, wj представляет вес, связанный с предсказанием классификатора. Существует два различных типа классификатора голосования. Они заключаются в следующем:

жесткий классификатор голосования.

мягкий классификатор голосования.

Классификатор жесткого голосования классифицирует входные данные на основе всех прогнозов, сделанных различными классификаторами. Когда голосование большинством голосов проводится на основе равных весов, выбирается режим прогнозируемой метки. Пусть, есть 3 классификатора: clf1, clf2, clf3. Для конкретных данных прогноз равен [1, 1, 0]. В случае, если веса, присвоенные классификаторам, равны, выбирается режим прогнозирования. Таким образом, режим [1, 1, 0] равен 1, и, следовательно, предсказанный класс для конкретной записи становится классом 1.

Классификатор мягкого голосования классифицирует входные данные на основе вероятностей всех прогнозов, сделанных различными классификаторами. Веса, применяемые к каждому классификатору. Давайте разберемся в этом на примере. Допустим, есть двоичные классификаторы clf1, clf2 и clf3. Для конкретной записи классификаторы делают следующие прогнозы в терминах вероятностей в пользу классов [0,1].

clf1 -> [0,2, 0,8], clf2 -> [0,1, 0,9], clf3 -> [0,8, 0,2]

При равных весах вероятности будут рассчитаны следующим образом:

класс 0 = 0.33*0.2 + 0.33*0.1 + 0.33*0.8 = 0.363,

класс 1 = 0.33*0.8 + 0.33*0.9 + 0.33*0.2 = 0.627.

Вероятность, предсказанная классификатором ансамбля, составит [36,3%, 62,7%]. Класс, скорее всего, будет классом 1, если пороговое значение равно 0,5. Вот как будет выглядеть классификатор мягкого голосования.

В машинном обучении измерение производительности является важнейшей задачей. Поэтому, когда дело доходит до проблемы классификации, рассчитываем на кривую AUC - ROC. ROC - это кривая вероятности, а AUC представляет степень или меру разделимости. Это говорит о том, насколько модель способна различать классы. Чем выше AUC, тем лучше модель предсказывает 0 классов как 0 и 1 класс как 1. По аналогии, чем выше AUC, тем лучше модель различает пациентов с заболеванием и без заболевания [1-3].

Для решения сформулированной задачи была реализована программа на языке python. Для сравнительного анализа результатов классификации была сформирована выборка, состоящая из 109 объектов, каждый из которых характеризуется 5 признаками (лейкоциты, нейтрофилы палочкоядерные, лимфоциты, общий билирубин, длительность оперативных вмешательств). Исходная выборка была разделена случайным образом на обучающую выборку и на тестовую выборку.

Были реализованы следующие методы машинного обучения для выявления наиболее точных и внесения их в метод ансамбль голосования.

```
The accuracy score for KNN is: 54.50000000000001%
The accuracy score for DTC is: 75.0%
The accuracy score for SVM is: 61.4%
The accuracy score for PPN is: 72.7%
The accuracy score for MLP is: 75.0%
The accuracy score for RandomForest is: 75.0%
```

Рис. 1. Точность методов классификации

Результаты, представленные на рисунке 1, показывают, что, наиболее точными методами стали деревья решений, многослойный пересептрон, случайный лес. Точность данных методов машинного обучения составила 75%. Следующим шагом, объединим данные классификаторы посредством классификатора мягкого голосования.

```
The accuracy score for method is: 77.3%
The f1 score for method is: 70.6%
The precision score for method is: 92.30000000000001%
The recall score for method is: 57.099999999999994%
```

Рис. 2. Точность классификатора
мягкого голосования

В результате работы классификатора мягкого голосования точность увеличилась до 77%. Результаты представлены на рисунке 2.

Библиографический список

1. Кобзарь, А.И. Прикладная математическая статистика / А.И. Кобзарь. – М.: Физматлит, 2006. – 816 с.

2. Чубукова, И.А. Data Mining: учебное пособие / И.А.Чубукова. – М.: Интернет–Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2006.

3. Айвазян С.А. Прикладная статистика: классификация и снижение размерности: справочное издание / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин / под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 608 с.

Кузнецов
Роман Николаевич
Кузнецова
Ольга Юрьевна
Пензенский государственный
университет,
г. Пенза, Россия

Kuznetsov R. N.
Kuznetsova O. Yu.
Penza State University,
Penza, Russia