УДК 004.85, 004.056

ВЕРИФИКАЦИЯ ДАННЫХ В УСЛОВИЯХ СИЛЬНОЙ НЕСБАЛАНСИРОВАННОСТИ КЛАССОВ

Колесник Д.П. (Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО")

Шуваев А.К. (Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО")

Федосенко М.Ю. (Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО")

Научный руководитель – к.т.н., доцент (квалификационная категория "ординарный доцент") Менщиков А.А. (Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО")

В работе рассматриваются основные методы классификации данных с помощью машинного обучения на несбалансированном наборе больших данных. Проведен обзор основных методов классификации и рассмотрены способы решения проблемы несбалансированности данных.

Введение. Алгоритмы машинного обучения широко применяются во многих сферах деятельности. Задача классификации обычно направлена на минимизацию количества ложных срабатываний. Например, в банковской сфере задача классификации может быть применена к обработке транзакций для выявления мошеннических действий или мошеннических схем. Однако при разработке самообучающейся модели возникает проблема несбалансированности данных, так как соотношение мошеннических транзакций к общему количеству транзакций слишком мало (по данным банка России доля мошеннических транзакций составила 0,0016 за 2017 год). При использовании стандартных методов классификации в такой ситуации часто возникает проблема, что при уменьшении общей ошибки классификатор полностью относит интересуемый класс к шуму. В таком случае возникает парадокс точности, когда показатели точности отражают только распределение базового класса. Таким образом, появляется потребность в решении проблемы несбалансированности.

Основная часть. Для решения поставленной задачи предлагается использовать несколько методов настройки классификаторов для несбалансированных данных. Метод уменьшения большего класса (Under Sampling) предлагает обучить модель на прецедентах первого класса и выбранных некоторым способом прецедентов 2 класса, увеличив таким образом соотношение нелегитимных транзакций к легитимным. С помощью метода увеличения класса (Over Sampling) возможно увеличить количество прецедентов меньшего класса, тем самым позволив модели обучаться на большем проценте нелегитимных транзакций. При работе на несбалансированных классах следует сменить главную метрику с точности на более подходящие метрики, такие как: Confusion Matrix, Precision, Recall, F1 Score (or F-score), Kappa (or Cohen’s kappa), ROC Curves. Таким образом, в рамках проводимой научно-исследовательской работы необходимо проанализировать основные методы машинного обучения для классификации данных, выбрать и реализовать подходящий метод решения в условиях сильной несбалансированности классов для реализации системы распознавания мошеннических транзакций.

Вывод. В работе был произведен обзор и сравненение уже существующих методов борьбы с сильной несбалансированностью классов. Было выявлено, что при реализации системы распознавания нелегитимных транзакций в банковской сфере главной проблемой является сильная несбалансированность классов в массиве данных. Требуется, чтобы модель машинного обучения в режиме реального времени с обучением без учителя в условиях сильной несбалансированности классов умела точно распознавать нелегитимные транзакции.

Колесник Д.П. (автор) Подпись

Федосенко М. Ю (автор) Подпись

Шуваев А.К. (автор) Подпись

Менщиков А.А. (научный руководитель) Подпись