Влияние дисбаланса классов в обучающей выборке на качество классификации для задачи определения литотипов по фотографиям полноразмерного керна

Д.О. Макиенко[1], И.А. Селезнев[1], И.В. Сафонов[1]

[1]Научно-исследовательский центр «Шлюмберже», Пудовкина 13, Москва, Россия, 119285

Аннотация. Использование методов машинного обучения широко распространено сегодня для решения прикладных задач в различных областях деятельности. Качество получаемых моделей обычно связано с качеством данных, используемых для обучения и настройки модели. На практике часто встречаются несбалансированные выборки данных, когда без учета этой особенности построить модель приемлемого качества не удается. Так, при построении классифицирующих и регрессионных моделей для фотографий полноразмерного керна, часто встречаются ситуации, когда превалирующий литотип представлен много большим количеством образцов, чем другие литотипы. При обучении модели такой дисбаланс затрудняет получение и обобщение информации о мало представленном (миноритарном) классе. Во-первых, в текущей выборке могут быть не отражены какие-либо свойства, характерные для данного литотипа. Во-вторых, свойства могут не извлекаться в силу влияния дисбаланса. В работе используется один из вариантов балансировки набора данных для обучения – увеличение числа примеров (oversampling) миноритарного класса. Рассматривается ряд примеров с разными характеристиками баланса данных, а также влияние размера исходной выборки на возможность коррекции дисбаланса.

1. Введение

Литологическое описание полноразмерного керна является трудоемким процессом. Привлечение фотографий полноразмерного керна для классификации горных пород и разметки соответствующих этим породам интервалов может существенно сократить время, требуемое для описания образцов. Работа по автоматизации описания горных пород с привлечением фотографий керна ведется и, как правило, подходы, предлагаемые для решения этой задачи, используют те или иные методы машинного обучения. При этом одним из основных типов данных для машинного обучения служат цветовые характеристики элементов изображения керна. [1,2] В этой работе также используются цветовые характеристики для построения предсказывающей модели. Не останавливаясь на деталях процесса, рассматривается важный момент, во многом определяющий качество классификации горных пород, а именно – влияние дисбаланса данных, на которых происходит обучение предсказывающей модели, и рассматривается одна из возможностей компенсировать такой дисбаланс.

Цель нашего исследования состоит в определении параметров выборки, существенно влияющих на качество предсказывающей модели, а также в оценке степени такого влияния.

Анализируя параметры выборок в широком диапазоне значений, мы хотим найти подмножества параметров, для которых применение выбранного метода корректировки дисбаланса оправдано и повышает качество предсказывающей модели.

2. Обзор методик для работы с несбалансированными данными

В задаче классификации предпочтительно, чтобы обучающие примеры были относительно равномерно распределены среди классов, так как некоторые классификаторы в равной мере учитывают ошибки обоих классов и в случае дисбаланса будут больше сконцентрированы на мажоритарном классе. Причиной такой работы классификаторов является то, что выявление характеристик мажоритарного класса внесет больше вклада в целевое значение (функционал качества или функцию ошибок), чем выявление характеристик миноритарного класса. При этом нередко в прикладных задачах встречаются несбалансированные наборы данных [3,4]. Не являются исключением и наборы данных для литологического описания керна. Дисбаланс классов может быть связан с различной встречаемостью пород. Для обучения модели на несбалансированных данных могут применяться следующие методики [3]:

1. Балансировка. Изменение соотношения классов в выборке путем увеличения числа экземпляров миноритарного класса (oversampling) или уменьшения числа экземпляров мажоритарного класса (undersampling).
2. Внесение корректировок в алгоритм обучения. Например, установление различных штрафов для классов в методе опорных векторов, изменение вероятностного порога отнесения примера к классу в деревьях.
3. Установление различной цены ошибок для классов. Цена ошибок может учитываться как при изменении соотношения классов в выборке, так и при внесении корректировок в алгоритм обучения.
4. Использование бустинга. Несколько классификаторов, корректирующих ошибки друг друга, могут повысить качество предсказаний модели на примерах миноритарного класса.

Использование таких методик для разных несбалансированных наборов данных рассматривалось в работах [4,5,6].

Для задачи литологического описания керна будет опробована одна из методик - балансировка выборки за счет увеличения числа примеров миноритарного класса. Для увеличения числа примеров миноритарного класса могут быть использованы следующие алгоритмы:

1. Random oversampling. Создаются копии случайно выбранных элементов миноритарного класса до достижения необходимого соотношения.
2. SMOTE (Synthetic Minority Oversampling Technique) [7]. Генерируются новые примеры за счет интерполяции примеров миноритарного класса - некоторого i-го примера и одного из его k-ближайших соседей. Существует несколько вариантов выбора i-го примера. Можно производить выбор случайным образом (Regular SMOTE), выбирать пример в зависимости от классов, которым принадлежат окружающие его примеры (Borderline SMOTE), в зависимости от построенных опорных векторов или от построенных кластеров.
3. ADASYN (Adaptive Synthetic) [8]. Работает аналогично методу SMOTE, но выбирает i-й пример миноритарного класса в зависимости от коэффициента r_i, который показывает долю примеров других классов вокруг i-го примера. Чем больше коэффициент r_i, тем больше примеров будет сгенерировано в окрестностях i-го примера.

3. Постановка эксперимента

В данной работе рассматривается задача классификации и разметки участков керна. Задача сегментации участков керна может быть решена путем выделения целевых литотипов на фоне остальных. Доля образцов, соответствующих литотипу, может меняться от скважины к скважине. Для исследования влияния дисбаланса на качество классификации были выбраны образцы, где исследуемый литотип представлен практически без дисбаланса, а также

разнообразно представлен на изображениях керна, отражая типичные цветовые особенности данного литотипа.

Для обучения модели и для исследования влияния дисбаланса используются интервалы, размеченные как «алевро-глинистая порода». Данные, полученные после обработки всех изображений керна, будут называться исходной выборкой. Для исследования влияния дисбаланса на качество классификации, из исходной выборки будут сформированы подвыборки разного размера, которые будут выступать в роли миноритарного класса. На каждой подвыборке будут обучены предсказывающие модели и исследована возможность компенсации дисбаланса.

В качестве исходных выборок используются 4 набора данных 2330:4075, 1165:2038, 583:1019, 292:510, где первое значение – число примеров миноритарного класса («алевро-глинистая порода»), второе – число примеров мажоритарного класса (другие литотипы). Для создания подвыборок с различными соотношениями классов и исследования влияния начальных соотношений на дальнейшее дополнение выборки, миноритарный класс уменьшается путем выбора случайным образом некоторого его подмножества, размером m. Значение m соответствует некоторому новому соотношению классов с долей примеров миноритарного класса p. Такие подвыборки могут обозначаться как $p(m)$. Чтобы уменьшить влияние случайного фактора на результаты классификации, для каждой подвыборки $p(m)$ примеры выбираются 10 раз и результаты по ним усредняются.

Различные уровни дисбаланса устанавливаются в обучающем наборе, тестовый при этом не меняется и имеет дисбаланс, присущий исходной выборке. Для оценки качества моделей используется кросс-валидация с разделением исходной выборки на 5 частей. Таким образом, обучающий набор состоит из 4/5 частей и до установления дисбаланса имеет размеры: 1864:3260, 932:1630, 466:815, 234:408. Тестирование производится 5 раз на каждой из частей и результаты усредняются.

Для балансировки обучающего набора используется методика увеличения числа примеров миноритарного класса SMOTE со случайным выбором примеров. После балансировки и обучения вычисляется точность классификации.

Для обучения классификатора используются два алгоритма: алгоритм для линейной классификации (логистическая регрессия) и алгоритм на основе деревьев (градиентный бустинг). Для иллюстраций результатов нашего исследования выбрана предсказывающая модель на базе логистической регрессии.

Для оценки качества моделей применяется F1-мера. При вычислении F1-меры используются две другие метрики, и она равна их среднему гармоническому:

$$Recall = \frac{TP}{TP + FN};\ Precision = \frac{TP}{TP + FP};\ F1 = \frac{2TP}{2TP + FP + FN}$$

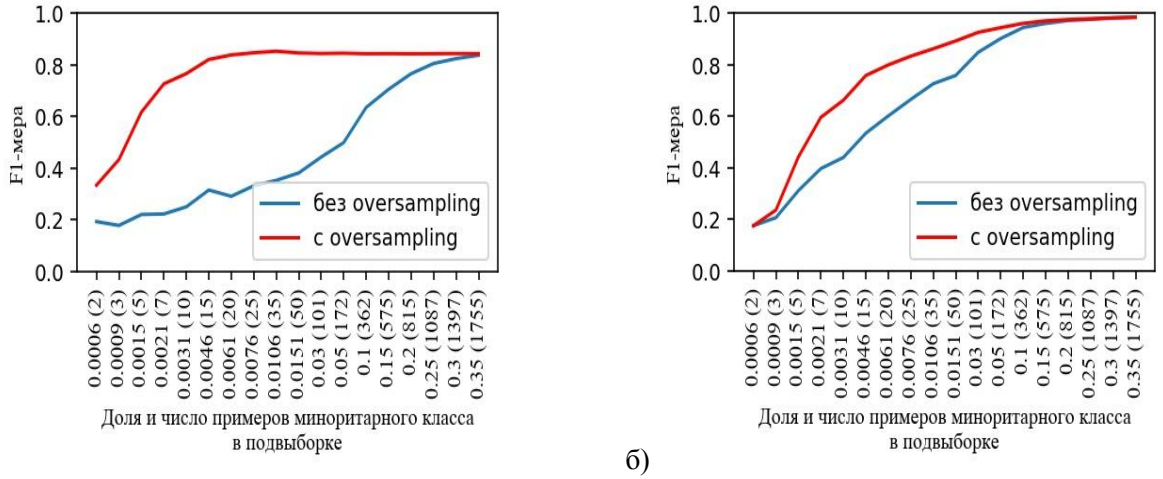
Здесь TP (True Positive), TN (True Negative) – количество верно предсказанных объектов положительного и отрицательного классов; FN (False Negative), FP (False Positive) – число объектов, неверно отнесенных к отрицательному и положительному классам. Положительным классом будет миноритарный класс, соответствующий алевро-глинистой породе, отрицательным – мажоритарный класс, соответствующий всем остальным породам.

4. Результаты

Качество моделей для определения литотипа «алевро-глинистая порода», обученных на несбалансированных подвыборках без применения oversampling, увеличивается при росте доли примеров миноритарного p класса и их количества m (рисунок 1). При этом качество выходит на приемлемый уровень лишь на подвыборках, на которых уровень дисбаланса совсем небольшой. Поэтому возникает необходимость корректировки дисбаланса, а также изучения влияния параметров p и m на работу классификатора. После применения oversampling для создания классов равных размеров качество моделей повышается.

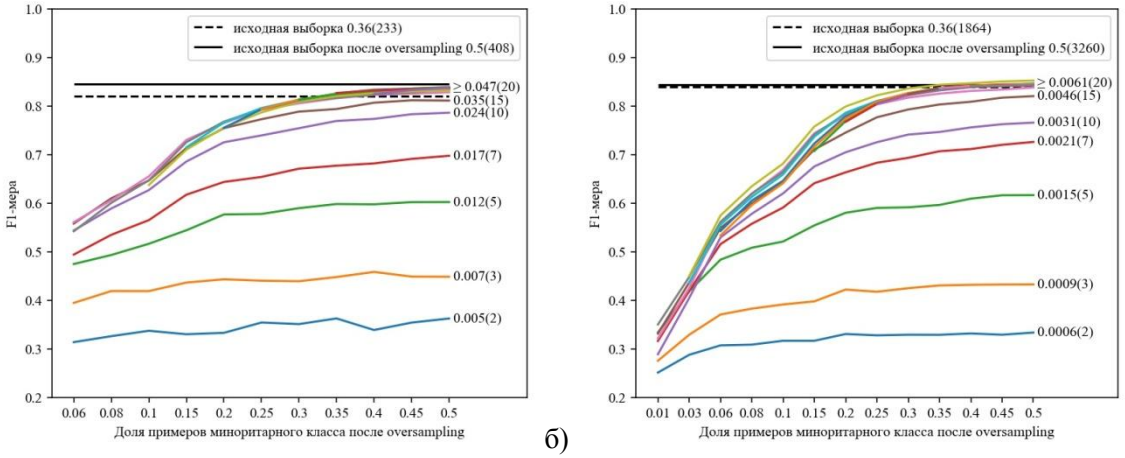
Для разных размеров мажоритарного класса одно и то же число примеров в миноритарном классе m соответствует разным уровням дисбаланса (разным долям p). В четырех обучающих

наборах путем уменьшения числа примеров миноритарного класса установлены разные уровни дисбаланса. После применения oversampling на качество модели влияет число примеров в миноритарном классе и не влияет уровень дисбаланса до искусственного увеличения класса. На рисунке 2 показаны графики зависимости F1-меры от доли примеров миноритарного класса в дополненных подвыборках для двух вариантов обучающих наборов. Графики для обучающих наборов 932:1630 и 466:815 не приведены ниже, так как они имеют схожее поведение. Справа от графиков указаны доля p и число m примеров миноритарного класса до дополнения.



а) б)

Рисунок 1. Качество классификации на несбалансированных подвыборках, полученных из обучающего набора 1864:3260 (а) логистическая регрессия, (б) градиентный бустинг.



а) б)

Рисунок 2. Оценка классификации для обучающих наборов (а) 234:408, (б) 1864:3260.

Исходя из полученных графиков, для выделения литотипа «алевро-глинистая порода» можно сделать следующие выводы:

1. Качество модели зависит от числа примеров в миноритарном классе до дополнения и не зависит от того, какому уровню дисбаланса соответствует это количество примеров.
2. Качество модели растет при увеличении числа примеров, представляющих миноритарный класс до дополнения.
3. Качество модели растет при увеличении доли примеров миноритарного класса за счет применения oversampling.
4. Существует некоторый порог для числа примеров m, при достижении которого может быть достигнуто качество исходной выборки, если применить oversampling.

При малом количестве примеров миноритарного класса случайный фактор при выборе этих примеров имеет существенное влияние на точность классификации. Если в обучающем наборе 1864:3260 установлен дисбаланс с долей примеров миноритарного класса p = 0.002 (m = 5), то при дополнении миноритарного класса до равенства с мажоритарным в среднем F1-мера равна

0.62, но максимальное отклонение от среднего для 10 вариантов таких подвыборок достигает 0.09. При увеличении числа примеров среднее значение F1-меры увеличивается, и уменьшается разброс (это верно не для всех m, но в общем такая тенденция сохраняется). Для подвыборки с параметрами p = 0.015 (m = 50) среднее значение F1-меры после дополнения до равных размеров классов равняется 0.85 и максимальное отклонение от среднего равняется 0.01. На рисунке 3 показаны графики зависимости F1-меры от степени дополнения выборки для 10 случайных извлечений примеров из исходной выборки при p = 0.002 (m = 5), p = 0.005 (m = 15) и p = 0.015 (m = 50).

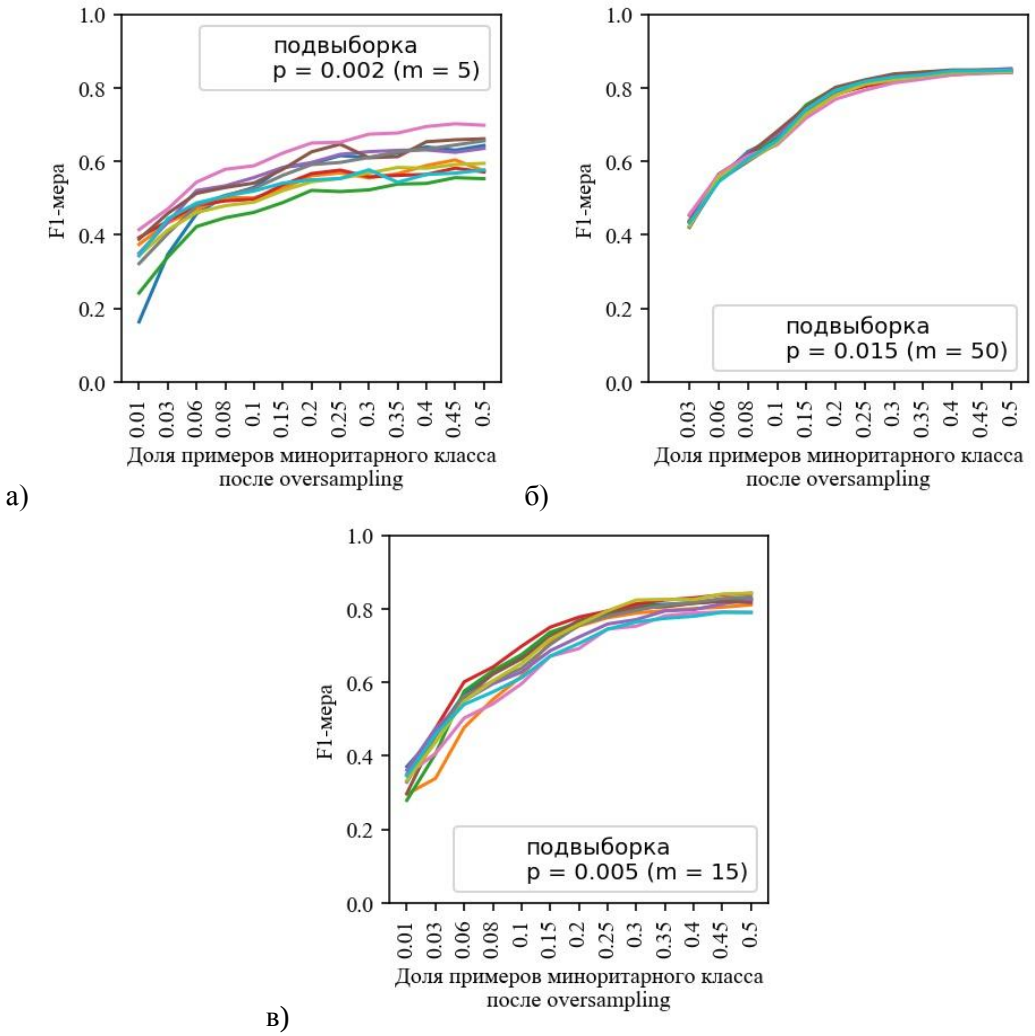


Рисунок 3. Графики, показывающие разброс F1-меры для 10 случайных вариантов подвыборок из обучающего набора 1864:3260 с параметрами (а) p = 0.002 (m = 5), (б) p = 0.005 (m = 15), (в) p = 0.015 (m = 50).

Для оценки схожести подвыборок, сбалансированных методом SMOTE до равного соотношения классов, с исходной выборкой использовались гистограммы и диаграммы рассеяния, построенные для одного и двух наиболее значимых признаков соответственно. Для 10 вариантов подвыборок с параметром m = 50, сбалансированных до равного соотношения классов, распределение признаков на гистограммах и диаграммах рассеяния визуально схоже с распределением исходной выборки (рисунок 4 (в), (г)). Для подвыборок с параметром m = 5, сбалансированных до равного соотношения классов, распределения признаков могут быть близкими к распределению исходной выборки, но в большинстве случаев имеют существенные различия (рисунок 4 (а), (б)).

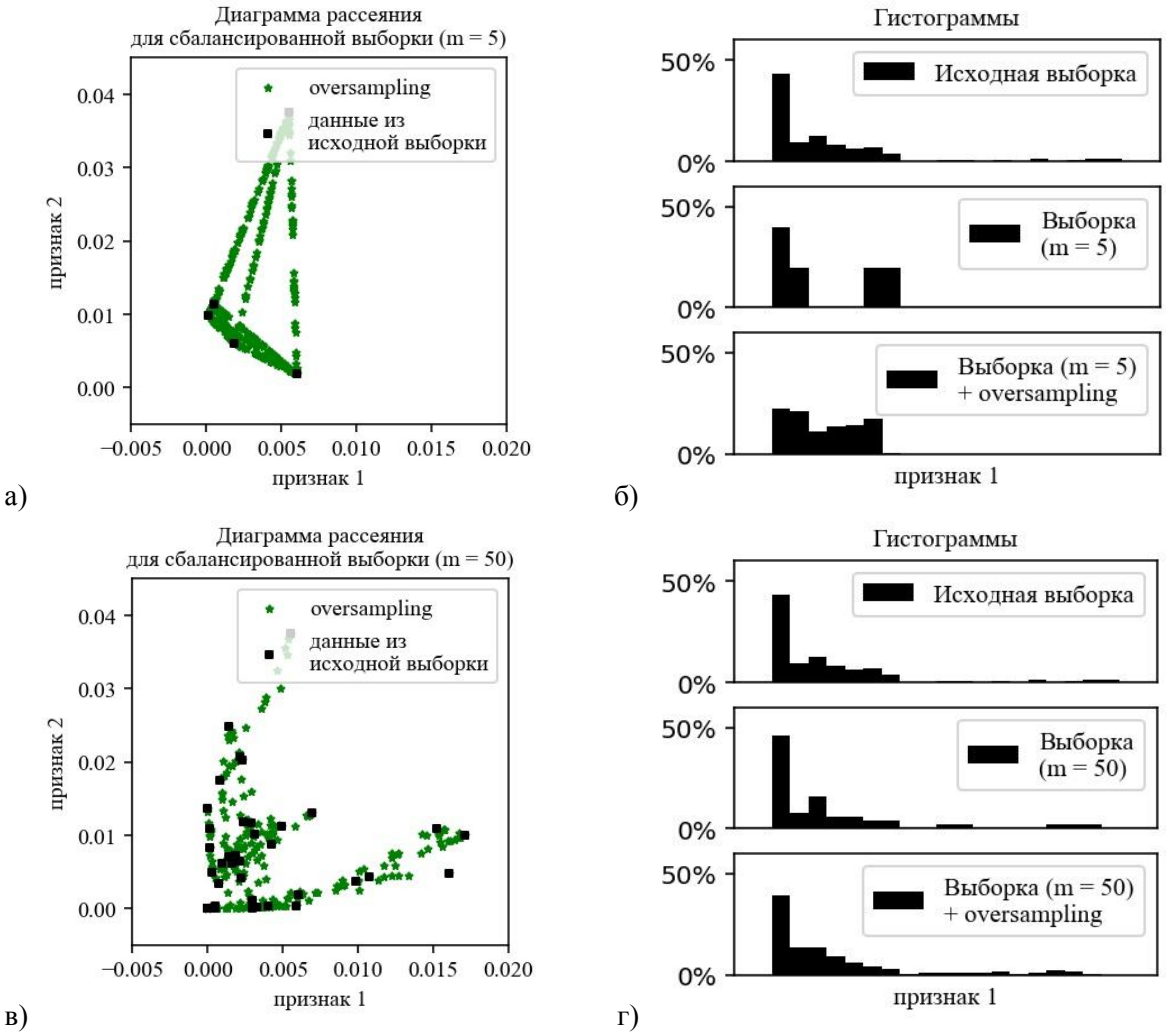


Рисунок 4. Диаграммы рассеяния и гистограммы для (а) - (б) сбалансированной подвыборки при $m = 5$, (в) - (г) сбалансированной подвыборки при $m = 50$. Подвыборки сбалансированы до равного размера классов.

5. Заключение

В работе рассмотрено влияние дисбаланса данных на качество классификации литотипов по фотографиям полноразмерного керна. Показано, что качество предсказывающих моделей, обученных на несбалансированных данных, сильно зависит от степени дисбаланса и часто неприемлемо, что вызывает необходимость повышения степени балансировки.

Показано, что величина дисбаланса, при которой можно получить предсказывающую модель близкую по качеству к модели, обученной на сбалансированной выборке, не постоянна и зависит от размера выборки данных доступных для обучения модели, а также от качества составляющих выборку образцов данных. Под качеством здесь понимается, насколько полно представленные образцы отражают особенности целевого литотипа.

Увеличение степени балансировки данных методом SMOTE привело к повышению качества определения целевого литотипа на примере выявления алевро-глинистой породы на фоне остальных представленных литотипов. Качество предсказывающей модели, близкое к качеству модели, построенной на всем имеющемся наборе данных, достигалось при таком числе образцов, которое позволяло восстановить распределение признаков всего набора данных с наименьшим влиянием случайного фактора при выборе примеров.

На рассмотренных примерах показано, что существует минимально допустимое количество образцов, слабо зависящее от размера всей выборки, при котором можно говорить

об воспроизводимом качестве обучения модели (с приемлемой дисперсией критерия качества). При уменьшении количества образцов доступных для обучения, дисперсия критерия качества модели возрастает.

6. Литература

[1] Baraboshkin, E.E. Deep convolutions for in-depth automated rock typing / E.E. Baraboshkin, L.S. Ismailova, D.M. Orlov, E.A. Zhukovskaya, G.A. Kalmykov, O.V. Khotylev, E.Y. Baraboshkin, D.A. Koroteev // Computers & Geosciences. – 2019. – Vol. 135. – P. 104330.

[2] Thomas, A. Automated lithology extraction from core photographs / A. Thomas, M. Rider, A. Curtis, A. MacArthur // First Break. – 2011. – Vol. 29(6). – P. 103-109.

[3] Sun, Y. Classification of imbalanced data: A review / Y. Sun, A.K. Wong, M.S. Kamel // International Journal of Pattern Recognition and Artificial Intelligence. – 2009. – Vol. 23(04). – P. 687-719.

[4] Паклин, Н.Б. Построение классификаторов на несбалансированных выборках на примере кредитного скоринга / Н.Б. Паклин, С.В. Уланов, С.В. Царьков // Искусственный интеллект. – 2010. – № 3. – С. 528-534.

[5] Seiffert, C. Building Useful Models from Imbalanced Data with Sampling and Boosting / C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano // FLAIRS conference. – 2008. – P. 306-311.

[6] Weiss, G.M. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? / G.M. Weiss, K. McCarthy, B. Zabar // Dmin. – 2007. – Vol. 7(35-41). – P. 24.

[7] Chawla, N.V. SMOTE: synthetic minority over-sampling technique / N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer // Journal of artificial intelligence research. – 2002. – Vol. 16 – P. 321-357.

[8] He, H. ADASYN: Adaptive synthetic sampling approach for imbalanced learning / H. He, Y. Bai, E.A. Garcia, S. Li // IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008. - P. 1322-1328.

The effect of the imbalanced training dataset on the quality of classification of lithotypes via whole core photos

D.O. Makienko[1], I.A. Seleznev[1], I.V. Safonov[1]

[1]Schlumberger Moscow Research, Pudovkina Street 13, Moscow, Russia, 119285

Abstract. Nowadays machine learning methods play an important role in many industries. Mainly, the effectiveness of the predicting models depends on the dataset. In practice, the imbalanced datasets are quite common. Frequently in the task of model creation for classification of lithotypes via whole core photos prevailing lithotype is represented by a much larger number of samples than other lithotypes. It is difficult to obtain good generalization in training for poorly represented classes. Firstly, some characteristics of a given minor lithotype may be absent. Secondly, some features of a minor class can be ignored due to imbalance. For the problem of speeding-up of the geological core description, we analyze the oversampling of a minor class to obtain the balanced dataset. Several examples with different data balances as well as the effect of the size of the original dataset on the effectiveness of correcting the imbalance are considered.