О методе повышения точности аннотирования изображений в краудсорсинге

О.Р. Нурмухаметов[1]
oleg.nurmuhametov@gmail.com

А.П. Бакланов[1,2,3]
baklanov@iiasa.ac.at

1 – ИММ УрО РАН (Екатеринбург) 2 – УрФУ (Екатеринбург)
3 – Международный институт прикладного системного анализа (Лаксенбург, Австрия)

Аннотация

Краудсорсинг представляет собой новый подход к решению задач, когда группа добровольцев заменяет экспертов. Последние результаты показывают, что краудсорсинг является эффективным инструментом для аннотирования больших массивов данных. Geo-Wiki является одним из успешных краудсорсинговых проектов. Основная цель проекта Geo-Wiki — улучшение глобальной карты земельного покрова путем применения краудсорсинга для распознавания образов. В данной работе исследуется методы повышения точности данных, собранных во время проведения игры The Cropland Capture (Geo-Wiki). В этой связи проведен анализ всех основных этапов краудсорсинговой кампании: обработка изображений и агрегация голосов. В ходе исследования используются методы компьютерного зрения и машинного обучения, которые позволили повысить оценку точности итогов голосования с 76% до 87%.

1 Введение

Изучение растительного покрова поверхности Земли принципиально важно для охраны природы и сохранения разнообразия видов, управления лесными и водными ресурсами, планирования городской и транспортной структуры, планирования мер по предотвращению природных бедствий и смягчению их последствий, а также для развития агропромышленного комплекса. В этой связи задача повышения качества глобальных карт растительного покрова представляется важной и актуальной. Современные подходы к решению данной задачи связаны с анализом больших массивов аэрокосмических снимков. Одним из успешных примеров в этой области считается международный проект Geo-Wiki [1], использующий технологию краудсорсинга.

Краудсорсинг (crowdsourcing) (см., напр. [2]) – это относительно новый подход к решению задач принятия решений (в том числе, в области анализа изображений), при котором окончательное решение принимается путем согласования индивидуальных мнений добровольцев, распределенных по всему миру. Поскольку волонтеры не всегда являются специалистами в заданной предметной области и вольны в любой момент изменить избранную ими стратегию оценивания (например, с добросовестной на «вредительскую»), собранные данные требуют дополнительного исследования на предмет достоверности.

In: A.A. Makhnev, S.F. Pravdin (eds.): Proceedings of the 47th International Youth School-conference "Modern Problems in Mathematics and its Applications", Yekaterinburg, Russia, 02-Feb-2016, published at http://ceur-ws.org

В работе проводится анализ краудсорсинговой кампании The Cropland Capture. В этой кампании участвовало 2 783 добровольца, которые на протяжении 6 месяцев оценивали 192 613 изображений на наличие пахотных земель. Игровой процесс описывается в статье [3]. Применяемый нами подход в целом следует работе [4], в которой предлагается использовать метод краудсорсинга для приближенного моделирования мнений экспертов. В рамках предлагаемого метода мнения добровольцев агрегировались по правилу простого большинства. Точность такой модели составила 76% [5]. Предложенный нами подход позволяет повысить точность до 87%.

В рамках проведенного анализа были использованы методы поиска копий изображений [6] и изображений низкого качества [7]. Применение этих методов к существующему набору данных позволило уменьшить шум, а также понизить размерность признакового пространства задачи и повысить статистическую значимость окончательных результатов кампании.

При агрегации голосов волонтеров использовались следующие алгоритмы машинного обучения: линейный дискриминантный анализ, Random Forest и AdaBoost. По результатам численных экспериментов наилучшим оказался метод линейного дискриминантного анализа.

2 Данные

Результаты игры были зафиксированы в двух таблицах. Первая таблица содержит детальную информацию об изображениях: *imgID* — уникальный идентификатор изображения; *link* — ссылка на изображение; *latitude* и *longitude* — гео-координата, отвечающая центру изображения; *zoom* — детализация изображения (значения: 300, 500, 1000 м). Табл. 1 демонстрирует фрагмент информации по изображениям.

Таблица 1: Структура данных в таблице изображений.

imgID	*link*	*latitude*	*longitude*	*zoom*
3009	`http://cg.tuwien.ac.at/~sturn/crop/img_-112.313_42.8792_1000.jpg`	42.8792	-112.313	1000
3010	`http://cg.tuwien.ac.at/~sturn/crop/img_-112.313_42.8792_500.jpg`	42.8792	-112.313	500
3011	`http://cg.tuwien.ac.at/~sturn/crop/img_-112.313_42.8792_300.jpg`	42.8792	-112.313	300
30015	`http://cg.tuwien.ac.at/~sturn/crop/img_87.8458_26.2958_500.jpg`	26.2958	87.8458	500
30016	`http://cg.tuwien.ac.at/~sturn/crop/img_87.8458_26.2958_300.jpg`	26.2958	87.8458	300

Все голоса волонтеров собраны во второй таблице: *ratingID* — уникальный идентификатор рейтинга; *imgID* — идентификатор изображения; *volunteerID* — идентификатор волонтера; *timestamp* — время, когда рейтинг был получен; *rating* — ответ волонтера. Возможные значения *rating* следующие: 0 (‘Maybe’), 1 (‘Yes’), −1 (‘No’). Табл. 2 содержит часть данных из таблицы рейтингов.

Таблица 2: Структура данных по протоколу голосования.

ratingID	*imgID*	*volunteerID*	*timestamp*	*rating*
75811	3009	178	2013-11-18 12:50:31	1
566299	3009	689	2013-12-03 08:10:38	0
641369	3009	1398	2013-12-03 17:10:39	−1
2191752	3009	295	2013-12-21 05:39:42	1
3210996	3009	295	2014-02-15 08:08:44	−1
3980868	3009	1365	2014-04-10 16:52:07	1

По окончании краудсорсинговой компании было собрано 4.6 млн голосов от 2 783 волонтеров. Протокол голосования был сконвертирован в матрицу рейтингов. Матрица состоит из рейтингов, которые волонтеры

(строки матрицы) присвоили изображениям (столбцы матрицы):

$$R = \left(r_{v,i}\right)_{v=1,i=1}^{|V|,|I|}, \tag{1}$$

где V — множество всех волонтеров;

I — множество всех изображений, которые получили хотя бы один голос;

$r_{v,i}$ — голос волонтера v по изображению i.

Из-за нечеткого определения, ответ *'Maybe'* оказалось сложно интерпретировать. Поэтому мы рассматриваем ответ *'Maybe'* как ситуацию, в которой пользователь не видел изображение; обе ситуации кодируются как 0. Если волонтер имеет несколько голосов для одного и того же изображения, тогда *используется только последний голос.*

3 Методология

3.1 Поиск копий изображений

Первичный анализ набора данных показал, что ссылки на изображения не уникальны. При группировке по значению в поле *link* существует несколько значений *imgID*, что означает, что одним и тем же изображениям соответствуют различные идентификаторы в базе данных. Если сгруппировать копии ссылок, новый набор данных будет содержать только изображения, соответствующие уникальным ссылкам. Однако следует отметить, что по различным ссылкам могут быть расположены одинаковые изображения, так как набор данных для краудсорсинговой компании был сформирован путем объединения изображений из различных источников [3]. Следовательно, практически одни и те же изображения могут быть представлены различными записями в базе данных. Поэтому мы проверили датасет на наличие изображений с копиями другим способом, предварительно загрузив все 192 613 изображений jpeg (размер 512×512 пикселей). Итоговый размер всех файлов составляет около 9 ГБ.

3.1.1 Сравнение бинарных файлов изображений

Побитовая операция сравнения файлов изображений является крайне неэффективной процедурой. Очевидно, что в данном случае сравнение на основе значений хешей значительно более эффективно, чем при сравнении пиксель за пикселем. Для расчета хеш-значений файлов были применены два варианта различных функций: MD5 [8] и SHA (Secure Hash Algorithm) [9]. Главное отличие функций — длина хеша, а значит и вариативность. При этом более длинное значение хеша требует больше времени для вычисления. MD5 формирует хеш длиной 128 бит (10^{36} уникальных значений) на базе бинарного файла. Работает очень быстро, коллизии очень редки, но возможны [10]. Чтобы быть уверенным в отсутствии коллизий с высокой долей вероятности (актуально при огромной коллекции изображений), следует использовать SHA-512. Этот метод формирует хеш длиной 512 бит (10^{153} уникальных значений хешей).

Для коллекции из 192 613 изображений применение SHA-512 и MD5 дало одинаковые результаты. Вычисление хеш-функций на этой коллекции позволило обнаружить 32 099 изображений, для которых хеш не уникален. Бинарные файлы дополнительно были проанализированы побитово, чтобы гарантировать отсутствие ошибочных совпадений из-за коллизий. Все файлы, у которых совпал хеш, также были эквивалентны при побитовом сравнении. Следует отметить, что группировка изображений по ссылкам позволила найти только 29 397 изображений. Все эти изображения полностью содержатся в множестве дублей при сравнении бинарных файлов.

3.1.2 Сравнение контента изображений

Совпадение бинарных файлов влечет полное совпадение изображений. При этом обратное не всегда справедливо. Для изображений на рис. 1 человек без специальных технических средств не сможет заметить разницу, но фактически файлы разные, а предложенные ранее методы дадут неверный результат. Для поиска похожих изображений среди изображений с уникальными ссылками и бинарными файлами используются методы компьютерного зрения.

Мы используем перцептивную хеш-функцию для выявления таких случаев. Пример таких функций — это aHash, dHash и pHash [6]. Мы обнаружили, что pHash работает гораздо лучше, чем более быстрые методы aHash и dHash. Методы aHash и dHash находят дубликаты изображений, но при этом для изображения на рис. 2 также являются одинаковыми.

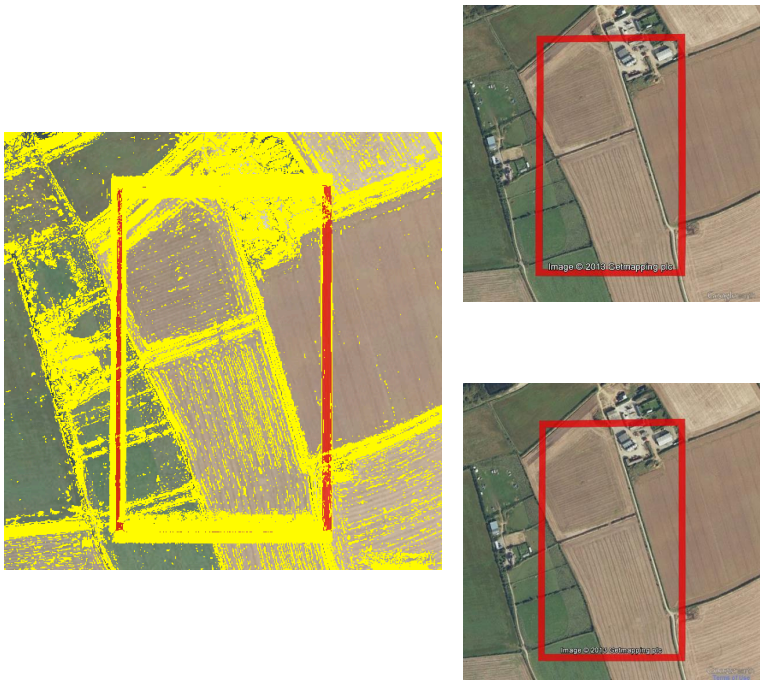

Рис. 1: Результат попиксельного сравнения двух внешне одинаковых изображений. Различия в изображениях для компьютера возникают из-за конвертации в разные форматы и изменения пропорций оригинала.

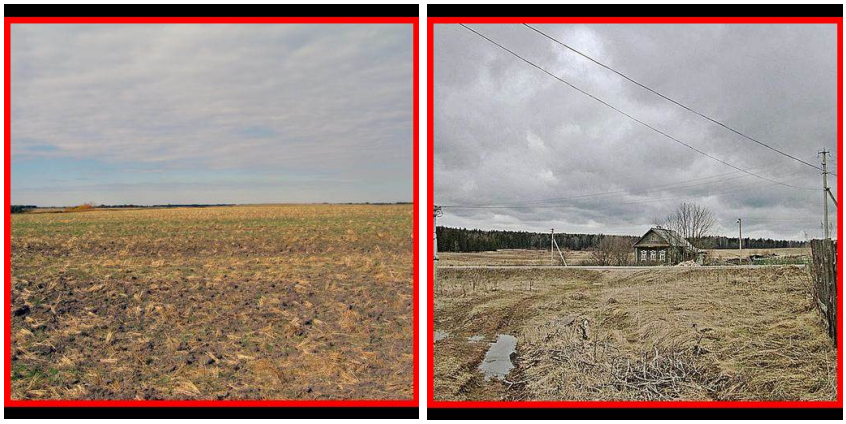

Рис. 2: Изображения, которые имеют одинаковое значение хеша aHash.

Отметим, что pHash позволяет сделать вывод о том, насколько два изображения различны. Для этого мы должны вначале вычислить хеш-значения изображений, а затем определить соответствующее расстояние Хемминга [11] для этих хешей. Чем больше расстояние, тем менее похожи друг на друга изображения. Нулевое расстояние означает, что это, скорее всего, одинаковые изображения (или вариации одного и того же изображения). Эта особенность выгодно отличает pHash от других хеш-функций.

Применение pHash для всего набора данных выявило множество из 39 000 изображений, среди которых только 8 300 являлись уникальными. Было проверено, что все 32 099 изображений, которые были получены предыдущим методом, также обнаруживаются методом сравнения хешей, рассчитанных на основе изображений. Метод сравнения изображений по контенту позволил обнаружить дополнительно более 7 тысяч неуникальных изображений. Для всех этих случаев результаты голосования были объединены. После объединения всех дублей изображений количество изображений сократилось с 192 613 до 170 041.

Если принять гипотезу мудрости толпы [12], то для принятия более точного решения по изображению необходимо собрать как можно больше голосов для каждого изображения. Объединение голосов по всем копиям изображения увеличивает статистически значимые эффекты и уменьшает размерность данных.

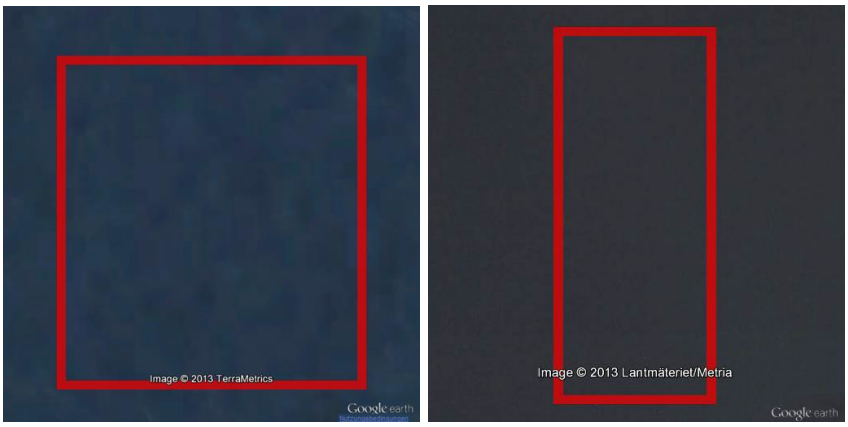

Рис. 3: Для данных изображений расстояние Хемминга для хешей pHash равно 5.

Кроме того, если обнаружение копий выполняется перед началом кампании, то происходит сокращение объема работы для волонтеров.

Таблица 3: Сравнение методов поиска дублей изображений.

Метод	*Кол-во копий*	*Кол-во уникальных*	*Затраченное время*
Текстовое сравнение ссылок (Links)	29 000	7 000	5 мин
Бинарное сравнение файлов (MD5/SHA)	31 700	7 930	15 мин
Визуальное сравнение изображений (pHash)	39 200	8 310	410 мин

3.2 Детектирование невозможных для распознавания изображений

Визуальный анализ изображений показал наличие неразличимых и размытых (несфокусированных) изображений. Как и следовало ожидать, эти изображения вызывали сомнения у волонтеров. Протокол голосования для таких изображений содержал большое количество голосов 'Maybe', количество голосов 'Yes'/'No' часто было практически одинаковым. Поэтому мы использовали автоматический метод поиска нечетких изображений, а именно *Blur Detection algorithm* [7].

Для каждого изображения был рассчитан уровень качества в диапазоне $[0, 1]$ таким образом, что 0 соответствует низкому качеству и 1 соответствует отличному качеству. Мы установили, что среди изображений с уровнем качества 0.2 и ниже, было всего несколько изображений, доступных для распознавания. Этот метод позволил нам обнаружить 2300 изображений настолько низкого качества, что даже эксперты не смогли бы дать правильный ответ. Заметим, что в дополнение к сильно размытым изображениям имеются изображения, которые также вызывают трудности у волонтеров. Эти изображения в основном содержат облака, тени от облаков или являются слишком темными.

Изображения со значением качества 0.70 достаточно хорошо распознаются человеком. На рис. 4 распознается текстура поля, фрагмент облака. Возможно, на результат повлияло наличие облака. Очень плохие изображения имеют качество 0.20 и ниже.

После консультации с экспертами было принято решение удалить все изображения качества ниже 0.2. Это уменьшило уровень шума и неопределенности в данных. После этого количество изображений уменьшилось с 170 041 до 161 752; окончательно $|I| = 161752$ и $|V| = 2783$.

Таблица 4: Статистика по качеству изображений в коллекции.

Уровень качества	*Количество*	*Процент*
1.00 (Отличные изображения)	74 000	38%
[0.00 - 0.20] (Крайне плохие изображения)	3 200	2%

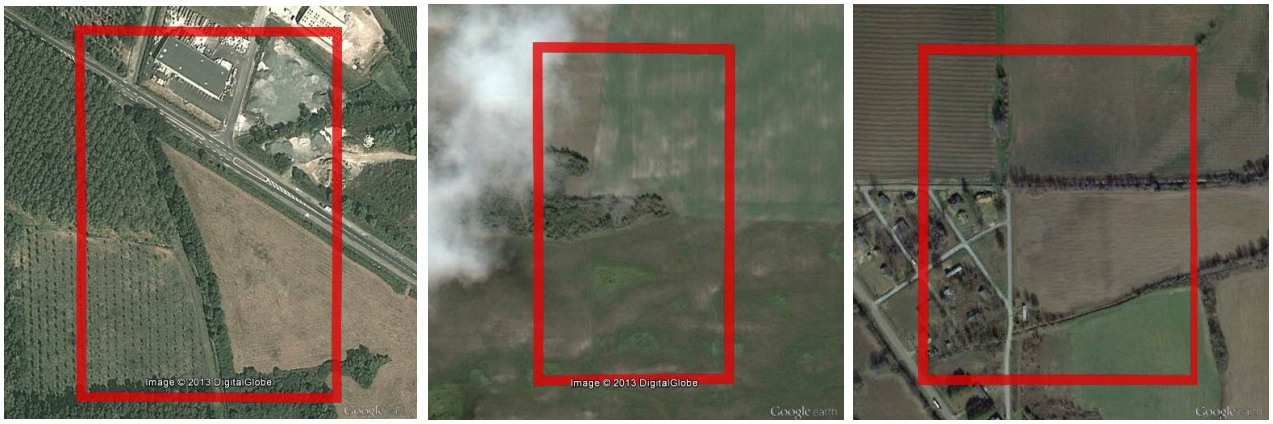

Рис. 4: Изображения хорошего качества: 1.00, 0.90, 0.70.

Рис. 5: Изображения низкого качества: 0.20, 0.10, 0.02.

4 Агрегация голосов

Чтобы воспользоваться стандартными алгоритмами машинного обучения, мы сначала применили *SVD (Singular Value Decomposition)* [13] ко всему набору данных. Отметим, что SVD очень широко применяется в машинном обучении [14] для автоматического выбора характеристик перед выполнением кластеризации и классификации. Исследование вариативности аппроксимации матрицы рейтингов помогло нам сделать выбор для числа признаков: 4, 14, 27.

Предоставленный экспертный набор данных для обучения и валидации алгоритмов состоял из 342 изображений. Экспертный набор данных был построен из двух групп изображений: простых и сложных для распознавания волонтерами [5]. По каждому изображению из набора было получено экспертное решение, как консенсус из двух независимых мнений специалистов в области дистанционного зондирования (remote sensing). На первичном наборе данных (до анализа изображений) точность правила большинства составила 76% [5]. Объединив копии и удалив сложные для распознавания изображения, мы получили новую версию экспертного набора данных из 194 уникальных изображения хорошего качества. Далее мы используем исключительно данную «урезанную» версию экспертного набора. Отметим, что при этом оценка точности правила большинства повысилась до 83%. Экспертная выборка была разделена в пропорции 70/30 на набор для обучения и набор для оценки точности (тестирования), который не участвовал в обучении.

Используя набор данных для обучения, мы варьировали параметры для алгоритмов:

- Random Forest [15]: мы подбирали число деревьев ([5, 10, 40, 100, 200, 300]), размер случайного подмножества, выбираемого на каждом шаге построения дерева ([20%, 40%, 60%, 80%, 100%]) в зависимости от количества признаков,
- AdaBoost [16]: мы подбирали итоговое количество оценивающих функций ([5, 10, 24, 50, 100, 200, 300]), максимальную глубину деревьев ([2, 10, 50, 100, None]), темп обучения [0.1, 0.3, 0.4, 0.6, 0.8, 1].

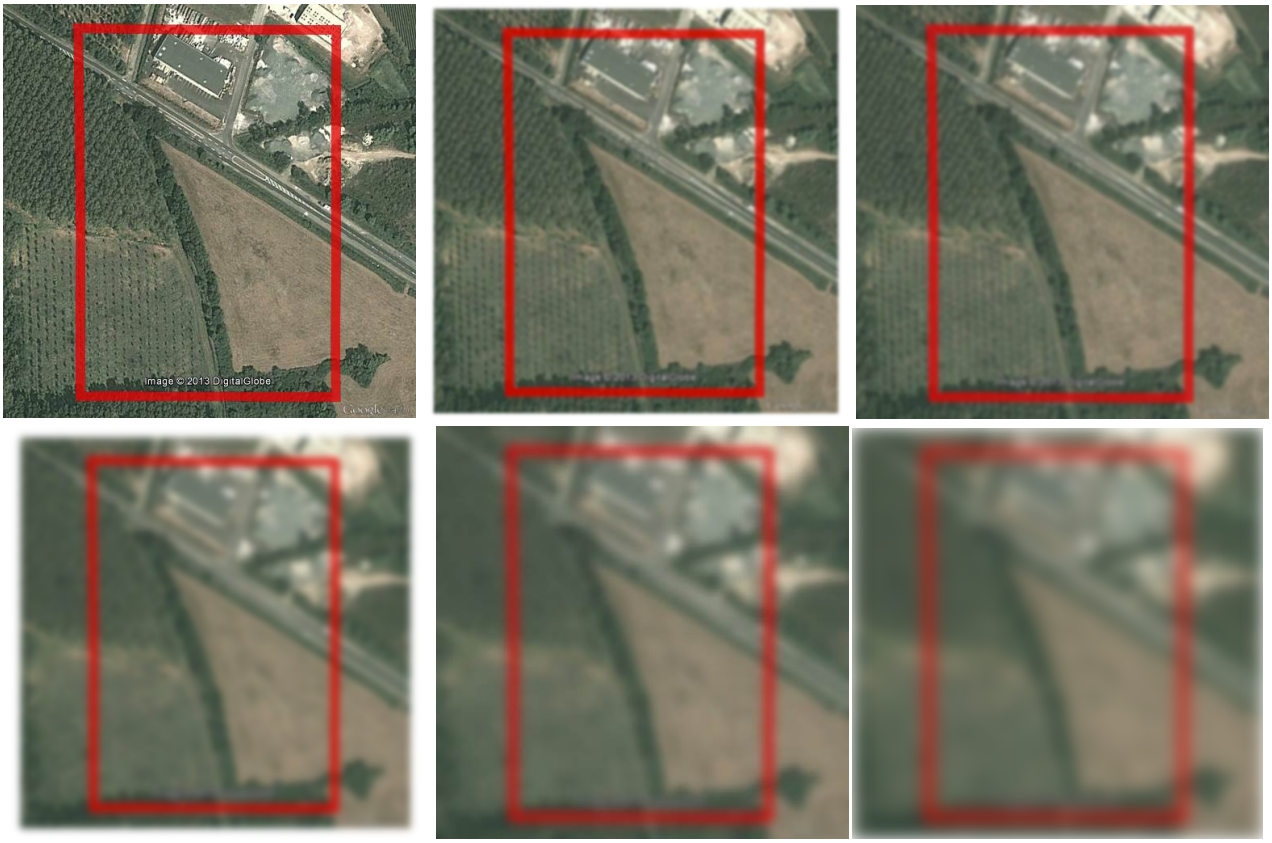

Рис. 6: Тестирование Blur Detection Algorithm. Изображение высокого качества при помощи фоторедактора было размыто. Качество изображений слева направо, сверху вниз: 1.00, 0.80, 0.54, 0.33, 0.21, 0.01.

Окончательный выбор параметров для AdaBoost и Random Forest был сделан на основе 10-кратной кросс-валидации на наборе данных для обучения. Для линейного дискриминантного анализа (Linear Discriminant Analysis, LDA) мы использовали параметры по умолчанию. Точность алгоритмов с подобранными параметрами была оценена при помощи тестового набора данных. В табл. 5 приведены итоговые результаты экспериментов. Отметим, что тестовое подмножество использовалось исключительно для оценки конечной точности алгоритмов обучения.

Таблица 5: Сравнение точности алгоритмов машинного обучения для агрегации данных.

Number of features	*Random Forest*	*LDA*	*AdaBoost*
4	83.90	83.88	83.82
14	77.61	**87.56**	84.36
27	75.72	84.84	85.24

В силу небольшой величины экспертной выборки можно допустить, что точность в табл. 5 является завышенной в силу эффекта переобучения. В дальнейшем мы планируем значительно увеличить экспертный набор данных, что позволит получить более надежные оценки точности.

5 Выводы и обсуждение

Дальнейшие исследования могут быть направлены на улучшение классификатора для поиска сложных и невозможных изображений для распознавания. Предложенный подход не позволяет находить снимки, которые следовало бы удалить: изображения, полностью покрытые облаками или тенью от них, ночные снимки.

Анализ проведенной кампании показал, что вовлеченность волонтеров на протяжении всей игры крайне неравномерная, есть большие всплески, а в некоторые дни волонтеры практически отсутствуют. Для эффективного проведения аналогичных компаний может быть важно научиться предсказывать количество пользователей и количество новых голосов.

Краудсорсинг — новый инструмент, имеющий широкие возможности. В работе мы использовали системный подход к проблеме качества полученной информации в краудсорсинге, который позволил выявить необходимые этапы при подготовке новых краудсорсинговых компаний по аннотированию изображений.

Благодарности

Работа выполнена при поддержке Российского научного фонда (проект 14-11-00109).

Список литературы

[1] S. Fritz, I. McCallum, C. Schill, C. Perger, R. Grillmayer, F. Achard, F. Kraxner, and M. Obersteiner. Geo-wiki.org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1(3):345–354, 2009.

[2] D. C. Brabham. *Crowdsourcing*. The MIT Press Essential Knowledge series, 2013.

[3] C. F. Salk, T. Sturn, L. See, S. Fritz, and C. Perger. Assessing quality of volunteer crowdsourcing contributions: lessons from the cropland capture game. *International Journal of Digital Earth*, 9(4):410–426, 2016.

[4] A. Comber, C. Brunsdon, L. See, S. Fritz, and I. McCallum. Comparing expert and non-expert conceptualisations of the land: an analysis of crowdsourced land cover data. In: *Spatial Information Theory*, 8116:243–260, Springer, 2013.

[5] C. F. Salk, T. Sturn, L. See, and S. Fritz. Limitations of majority agreement in crowdsourced image interpretation. *Transactions in GIS*, 2016.

[6] C. Zauner. *Implementation and benchmarking of perceptual image hash functions*. PhD thesis, 2010.

[7] H. Tong, M. Li, H. Zhang, and C. Zhang. Blur detection for digital images using wavelet transform. In: *IEEE International Conference on Multimedia and Expo, 2004 (ICME'04)*, 1:17–20, IEEE, 2004.

[8] R. Rivest. The md5 message-digest algorithm. 1992.

[9] D. Eastlake and P. Jones. Us secure hash algorithm 1 (sha1). 2001.

[10] X. Wang and H. Yu. How to break md5 and other hash functions. In: *Advances in Cryptology–EUROCRYPT 2005*, 19–35, Springer, 2005.

[11] R. W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[12] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[13] G. E. Forsythe and C. B. Moler. *Computer solution of linear algebraic systems*, vol. 7. Prentice-Hall Englewood Cliffs, NJ, 1967.

[14] P. Harrington. *Machine learning in action*. Manning, 2012.

[15] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In: *Computational learning theory*, 904:23–37, Springer, 1995.

A method for increasing the accuracy of image annotating in crowdsourcing

Oleg R. Nurmukhametov[1], *Artem P. Baklanov*[1,2,3]
1 – Krasovskii Institute of Mathematics and Mechanics (Yekaterinburg, Russia)
2 – Ural Federal University (Yekaterinburg, Russia)
3 – International Institute for Applied Systems Analysis (Laxenburg, Austria)

Keywords: crowdsourcing; Geo-Wiki; data quality; image classification.

Crowdsourcing is a new approach to solve tasks when a group of volunteers replaces experts. Recent results show that crowdsourcing is an efficient tool for annotating large datasets. Geo-Wiki is an example of successful citizen science projects. The goal of Geo-Wiki project is to improve a global land cover map by applying crowdsourcing for image recognition. In our research, we investigate methods for increasing reliability of data collected during The Cropland Capture Game (Geo-Wiki). In this regard, we performed analysis of all main steps of the crowdsourcing campaign: image processing and aggregation of collected votes. During the research, we used methods of Computer Vision and Machine Learning. This allowed us to increase accuracy of the aggregated votes from 76% to 87%.