МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное

образовательное учреждение высшего образования

«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет информатики и прикладной математики

Кафедра прикладной математики и экономико-математических методов

**КУРСОВАЯ РАБОТА**

по дисциплине:

**«Методы анализа данных»**

Тема: Реализация алгоритмов агрегации оценок исполнителей на краудсорсинговой платформе «Яндекс. Толока» для повышения качества разметки данных

Направление: 01.03.02 Прикладная математика и информатика

Студент: Широков Александр Анатольевич

Группа: ПМ1701 Подпись:

Проверил: Леора Светлана Николаевна

Должность: к.м.-ф.н., доцент

Оценка: Дата:

Подпись:

Санкт-Петербург

2020

**Оглавление**

[ВВЕДЕНИЕ 3](#_Toc59319942)

[1. ИСХОДНЫЕ ДАННЫЕ 4](#_Toc59319943)

[1.1. Постановка задачи 5](#_Toc59319944)

[1.2. Предварительный анализ данных 6](#_Toc59319945)

[1.3. Ошибка комитета большинства 9](#_Toc59319946)

[2. АЛГОРИТМЫ АГРЕГАЦИИ ОЦЕНОК ИСПОЛНИТЕЛЕЙ 11](#_Toc59319947)

[2.1. Простейшие методы агрегации данных 11](#_Toc59319948)

[2.1.1. Голос большинства (MV) 11](#_Toc59319949)

[2.1.2. Взвешенный голос большинства 11](#_Toc59319950)

[2.1.3. Методы оценки вероятности предоставления верного ответа 13](#_Toc59319951)

[2.2. Продвинутые методы агрегации 15](#_Toc59319952)

[2.2.1. Модель скрытой оценки 15](#_Toc59319953)

[2.2.2. Dawid & Skene алгоритм (DS) 18](#_Toc59319954)

[2.2.3. Generative Model of Labels, Abilities and Difficulties (GLAD) 19](#_Toc59319955)

[3. СРАВНЕНИЕ АЛГОРИТМОВ АГРЕГАЦИИ ОЦЕНОК ИСПОЛНИТЕЛЕЙ 22](#_Toc59319956)

[3.1. Сравнение алгоритмов агрегации по точности верных ответов на эталонных ответах 22](#_Toc59319957)

[3.2. Сравнение алгоритмов агрегации по времени работы 24](#_Toc59319958)

[Заключение 26](#_Toc59319959)

[Список ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 27](#_Toc59319960)

**ВВЕДЕНИЕ**

Толока – крупнейший источник размеченных людьми данных для задач машинного обучения.

Для любых исследований и экспериментов, связанных с машинным обучением, необходимы большие объёмы качественных данных. Раньше многие компании собирали такие оценки исключительно с помощью специально обученных сотрудников – асессоров. Но со временем задач в области машинного обучения стало слишком много, а сами задачи в массе своей перестали требовать особых знаний и опыта – так и появился спрос на помощь толпы (crowd). Так как самостоятельно найти большое количество случайных исполнителей и работать с ними не каждому под силу, то появились краудсорсинговые платформы, решающие эту проблему. Люди за плату выполняют задачи по различным темам: распознавание объектов и выделение областей на изображении, модерация изображений и видео, транскрипция аудио и многое другое.

Основным недостатком такого подхода является низкий уровень качества выполненных заданий. Тем не менее, исследования показывают, что польза от использования краудсорсинга есть, и результат, как правило, достигается при агрегации ответов нескольких исполнителей на одном и том же примере.

Целью данной курсовой работы является знакомство c подходами к агрегации меток исполнителей и реализация соответствующих алгоритмов агрегации оценок исполнителей на краудсорсинговой платформе «Яндекс. Толока». В данной работе рассмотрены различные методы агрегации меток, учитывающие различные факторы, такие как надёжность исполнителя, сложность задания или комбинация двух факторов сразу, а также различные продвинутые генеративные процессы, позволяющие выводить метки исполнителей с помощью идеи максимизации правдоподобия выборки в предположении, что ошибки совершаются исполнителями независимо.

Результатами курсовой работы являются сравнение методов агрегации по общей точности верных ответов и времени работы.

**1. ИСХОДНЫЕ ДАННЫЕ**

В качестве исходных данных были использованы два открытых набора данных, позволяющих исследовать различные методы контроля качества в краудсорсинге: Toloka Aggregation Relevance 2 (в дальнейшем AR2) и Toloka Aggregation Relevance 5 (в дальнейшем AR5).

Данные собирались следующем образом. Каждому исполнителю предоставлялся запрос, который ввёл пользователь для нахождения информации, а также результат выполнения данного запроса. В задаче бинарной классификации (AR2) исполнителю необходимо было определить является ли запрос релевантным или нет. В задаче оценки качества запроса по пяти градациям (AR5) исполнителю необходимо было оценить качество ответа на запрос по пятибалльной шкале – от нерелевантного (1) до витального (5). Пример задания, которые необходимо было решать исполнителю, представлен на рисунке 1 ниже.



Рисунок 1 – Пример решаемой исполнителем задачи

Из анонимизированных оценок исполнителей составлялись наборы данных AR2 и AR5. Основной задачей является по данным шумным оценкам исполнителей для каждого сайта определить его истинную метку, чтобы модели, обучающиеся на данных о качестве запросов, могли быть более точными и эффективными.

Качество моделей агрегации меток исполнителей оцениваются с помощью эталонных ответов – голденсетов — это наборы заданий, для которых известен правильный ответ. Эти данные позволяют понять, как мнение отдельных исполнителей влияет на качество итоговой оценки, какие методы агрегации результатов лучше использовать и сколько мнений нужно собрать, чтобы получить достоверный ответ.



Рисунок 2 – Пример эталонного ответа для задания

На рисунке 2 изображён пример эталонного ответа. В задаче бинарной классификации (AR2) правильная метка для сайта является единственной, а в задаче многоклассовой классификации (AR5) ввиду сложности оценки заданий для одного сайта может быть несколько правильных ответов.

**1.1. Постановка задачи**

Пусть имеется сайтов (), каждый из которых принадлежит одному из классов (. Имеется исполнителей (), каждый из которых некоторым сайтам поставил в соответствие определённый класс ( – исполнители, разметившие объект ):

Цель: имея ответы исполнителей для каждого примера предсказать истинную метку . Предполагается, что данная метка существует и единственна.

**1.2. Предварительный анализ данных**

Был проведён предварительный анализ данных, на которых тестировались различные алгоритмы агрегации.

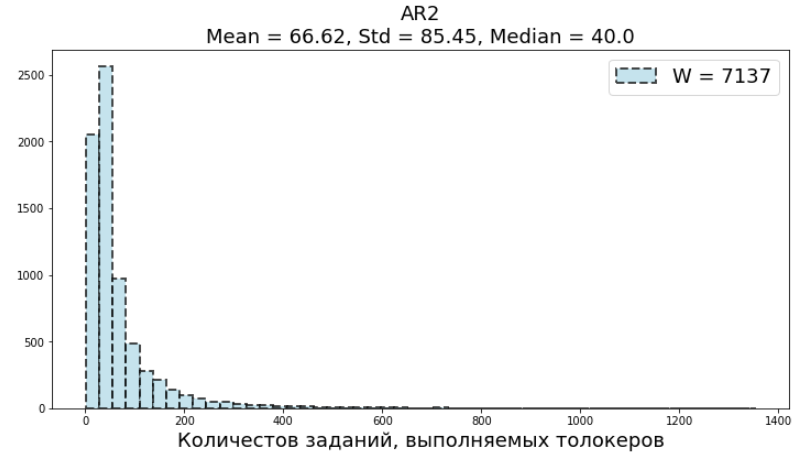


Рисунок 3 – Распределение количества заданий, выполняемых исполнителями на наборе данных AR2

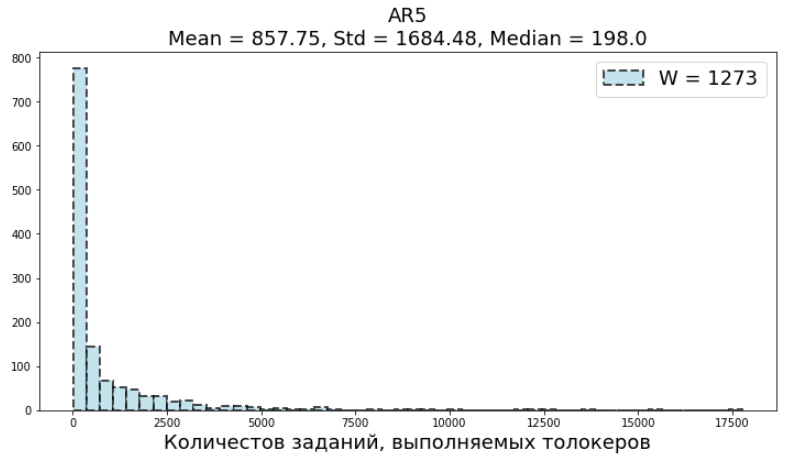


Рисунок 4 - Распределение количества заданий, выполняемых исполнителями на наборе данных AR5

На рисунках 3 и 4 представлено распределение количества заданий, выполняемых исполнителями на наборах данных AR2 и AR5 соответственно. Видно, что с первым набором данных работали намного больше исполнителей, чем на втором, из-за этого люди выполняли намного меньше заданий, чем на втором наборе данных.

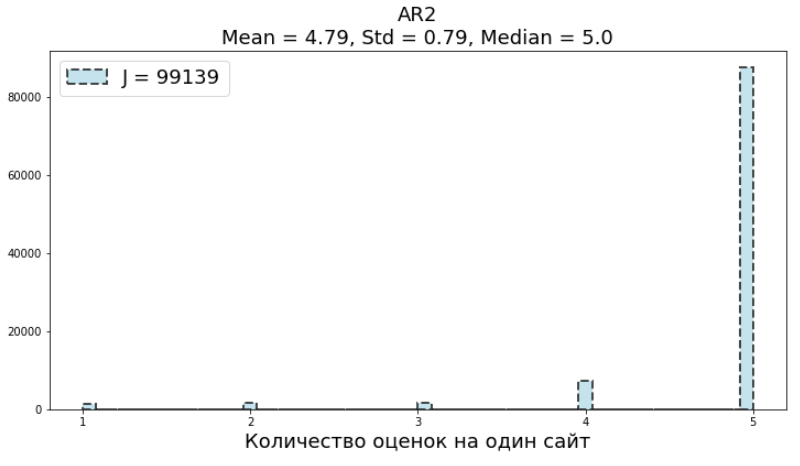


Рисунок 5 – Количество оценок на один сайт на наборе данных AR2

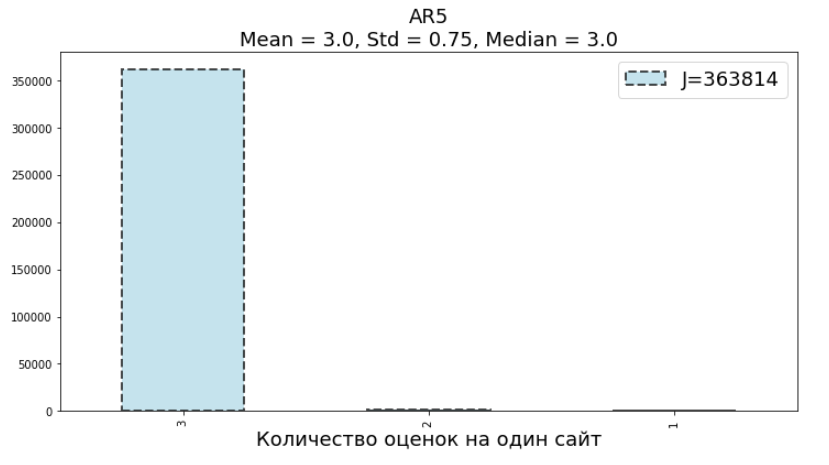


Рисунок 6 – Количество оценок на один сайт на наборе данных AR5

На рисунках 5 и 6 видно, что в большинстве своём на каждый сайт давалось по 5 оценок на наборе данных AR2 и по 3 оценки на наборе данных AR5.

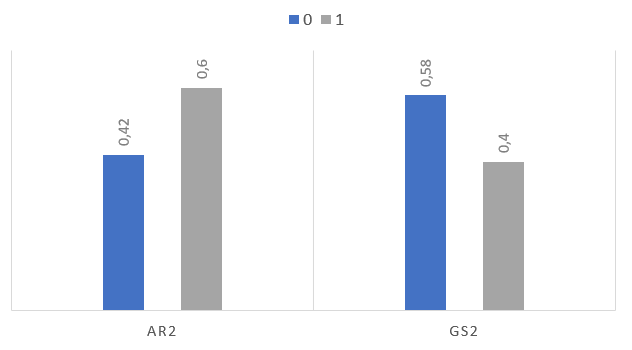


Рисунок 7 – Распределение классов на наборе данных AR2

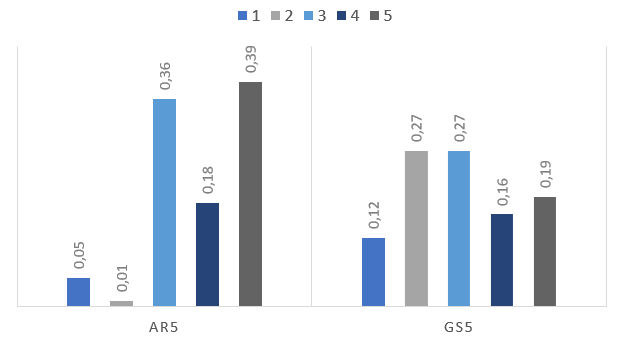


Рисунок 8 – Распределение классов на наборе данных AR5

На рисунках 7 и 8 отображено распределение классов на наборах данных AR2 и AR5 и на эталонных данных. Результаты показывают, что исполнители являются менее придирчивыми и более лояльными при оценке степени релевантности сайтов, из чего можно сделать вывод, что в любом случае из-за человеческого фактора агрегированные оценки исполнителей не могут быть абсолютно точными. Определим необходимое количество работников на одно задание, чтобы ошибка комитета большинства была наименьшей. С точки зрения теории вероятностей на это есть определённый ответ.

**1.3. Ошибка комитета большинства**

Пусть – количество работников, которые дали ответ на определённое задание, а – средняя точность ответов работников (*accuracy*).

**Теорема 1**. [5] Математическое ожидание вероятности того, что как минимум работников выдадут правильный результат при условии, что работники отвечают независимо:

Данное математическое ожидание можно оценить с помощью границы Чернова:

Обозначим за – необходимую точность результатов агрегации [1]. Тогда для заданной точности и средней точности работников , число работников на одно задание, гарантирующее с точки зрения теории вероятностей, что ожидаемая точность оценок агрегации будет не меньше необходимой точности , равно:

С помощью данного неравенства краудсорсинговые компании могут решать важную проблему – сколько мнений нужно собрать, чтобы получить достоверный ответ. Чем меньшее количество работников придётся привлечь на одно задание, тем меньшее количество денег придётся потратить на разметку данных. Поставим себя на место заказчиков и попытаемся оценить на основании наборов данных необходимое количество работников на одно задание. Среднюю точность верных ответов будем оценивать на эталонных ответах.

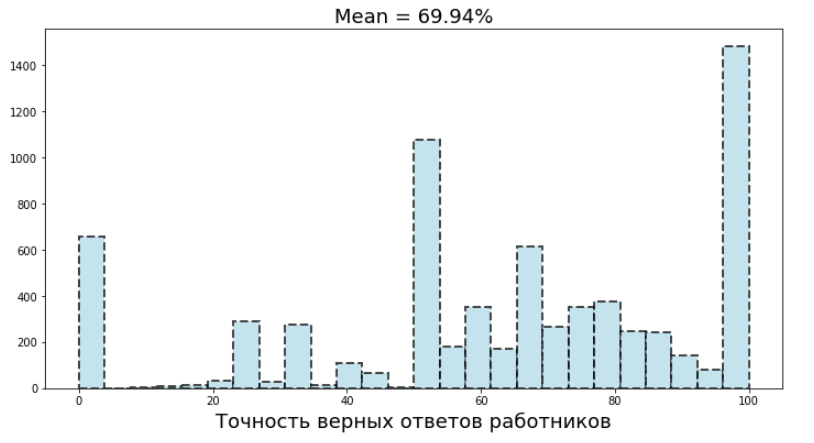


Рисунок 9 – Распределение точности верных ответов работников на наборе данных AR2, оценённых с помощью эталонных ответов

Видно, что средняя точность работников на данном виде заданий примерно 70 процентов, а для набора данных AR5 – примерно 85 процентов. Для необходимой точности возможно определить количество людей на одно задание с помощью рассмотренного неравенства.

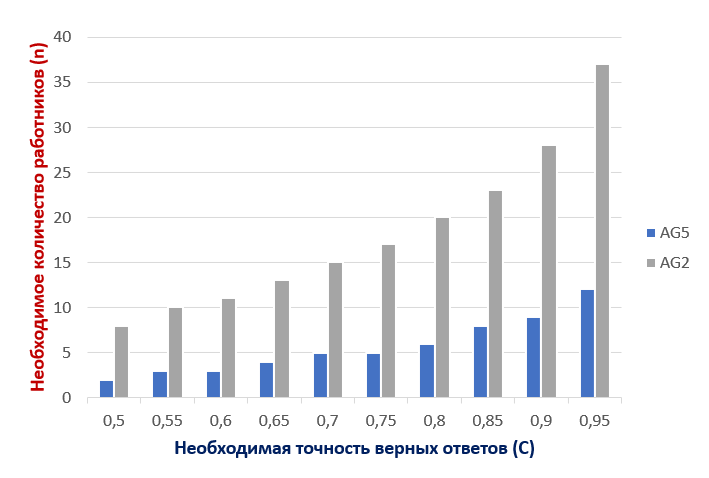


Рисунок 10 – Зависимость необходимой точности верных ответов от количества работников на задание.

Данное неравенство позволяет оценивать необходимое количество работников на задание. В случае, если квалифицированных работников не найти и приходится работать с имеющимися данным, существуют различные алгоритмы агрегации данных.

**2. АЛГОРИТМЫ АГРЕГАЦИИ ОЦЕНОК ИСПОЛНИТЕЛЕЙ**

**2.1. Простейшие методы агрегации данных**

**2.1.1. Голос большинства (MV)**

Пусть даны шумные оценки исполнителей и необходимо найти истинную оценку исполнителей . Очевидным решением является голосование большинства: для каждого примера будем выбирать метку, которая встречается чаще всего для данного примера. Этот способ называют *мажоритарным голосованием* (Majority Vote, MV):

Данный метод хоть и является простым и понятным, однако на практике является очень сильным базовым решением (baseline) и результат точности агрегации с помощью данного метода берётся за нижнюю оценку качества и результаты более продвинутых методов непременно сравнивают с результатом метода голоса большинства.

Однако, как уже было сказано, краудсорсингу присущ шум, и метки получаются от людей с разным уровнем компетентности и опыта – в данном же методе все исполнители и объекты считаются одинаковыми. Данный недостаток учтён в следующем методе агрегации.

**2.1.2. Взвешенный голос большинства**

Логичным кажется модифицировать метод голоса большинства, добавив вес каждому исполнителю как уровень его надёжности, отражающий вероятность предоставить правильную метку для произвольного примера, а затем проводить голосование с весами.

Данный метод порождает некоторый класс взвешенных методов агрегации. Данный класс включает в себя наиболее известных алгоритма, которые были рассмотрены и реализованы в курсовой работе:

1. Взвешенный голос большинства (Weighted Majority Vote, WMV)
2. Взвешенное среднее (Weighted Mean Aggregation, WMA)
3. Взвешенный голос успешных исполнителей (Weighted Majority Vote Threshold) – в данном методе учитываются лишь самые надёжные исполнители, например, вероятность предоставления правильного ответа больше определённого заранее заданного порога

Данные методы учитывают компетентность исполнителей и на практике дают значительное улучшение точности агрегации верных ответов.

В работе [4] доказано, что при известных – вероятностях предоставления правильного ответа оптимальным является байесовское голосование, где вероятность класса пропорциональна его правдоподобию:

Этот метод был назван Bayes Vote (BV) и в качестве истинной метки выдаётся следующее значение:

Данный алгоритм был так же реализован в данной курсовой работе и он действительно даёт значительный прирост в качестве. Как было показано, оценка надёжности исполнителя является основной задачей. Для проставления весов в данной задаче использовались эталонные ответы на задания и, оказывается, что от выбора метода оценки вероятности предоставления верного ответа каждым исполнителем также может зависеть точность агрегации.

**2.1.3. Методы оценки вероятности предоставления верного ответа**

Как уже было сказано, веса проставляются на основании известных эталонных ответов для заданий. В данной курсовой работе были рассмотрены способа оценок вероятностей предоставления верного ответа, которые будут подробно описаны ниже.

1. Первый метод определения вероятности проставления верного ответа каждого исполнителя рассматривает лишь точность верных ответов исполнителей на тестовом множестве и равно отношению количества верных ответов к общему числу проставленных ответов пользователем (W-метод):
2. Второй метод определения вероятности проставления верного ответа каждого исполнителя использует информацию о точности верных ответов работника, а также сложности заданий, с которыми работал данный исполнитель. Точность верных ответов исполнителей определяется также, как и в первом способе, а сложность заданий определяется как среднее арифметическое точности верных ответов других ответов работников на каждое из заданий. Чтобы перевести данные значения на язык вероятностей была использована логистическая функция, почему метод и получил своё название – логистическая сложность изображения (Difficulty of Images Sigmoid, DWS):
3. Третий метод был придуман и реализован автором курсовой работы. Он называется «Индикаторная сложность изображения» (Indicative Difficulty of Images) и задаётся следующей формулой:

В основе данного метода определения вероятности проставления верного ответа каждого исполнителя лежит следующая идея: оценивается точность исполнителя на эталонных заданиях и, если точность его верных ответов больше какого-то порогового значения (например ), то точность работы исполнителя является допустимой и неважно, с какими по сложности заданиями работал данный человек – исполнитель справляется с ними хорошо. Если же точность работника низкая, то осуществляется попытка оправдать исполнителя, посмотрев на сложность изображений, с которыми исполнитель работал. Если же и другие исполнители не могли хорошо ответить на данные задания, то мы «дотягиваем» сложность исполнителя до , иначе оправдания отсутствуют и вероятность предоставления правильного ответа данным исполнителем понижается ещё больше.

Данные метода были реализованы в данной курсовой работе и, как выяснилось, от выбора метода определения вероятности проставления верного ответа каждого исполнителя зависит общая точность результатов агрегации. Рассмотренные методы демонстрируют высокое качество результатов агрегации. Единственным недостатком является то, что данные веса берутся из информации о эталонных оценках, а ведь данных оценок может и не быть. Перейдём к продвинутым методам агрегации, учитывающие данные недостатки и предоставляющие новый способ решения задачи агрегации меток исполнителей.

**2.2. Продвинутые методы агрегации**

**2.2.1. Модель скрытой оценки**

Модели скрытой оценки предоставляют совершенно иной взгляд на то, как оценки исполнителей появляются в системе. Данные продвинутые методы позволяют не только находить агрегированные истинные оценки , но использовать параметризацию надёжности исполнителей и сложности объекта **.**

В более продвинутых методах мы предполагаем, что те оценки, которые наблюдаются от исполнителей являются случайными переменными**,** которые появляются согласно некоторому процессу. Для компактного изображения генеративных процессов, согласно которым появляются шумные оценки, используются графические модели, показанные на рисунке ниже.

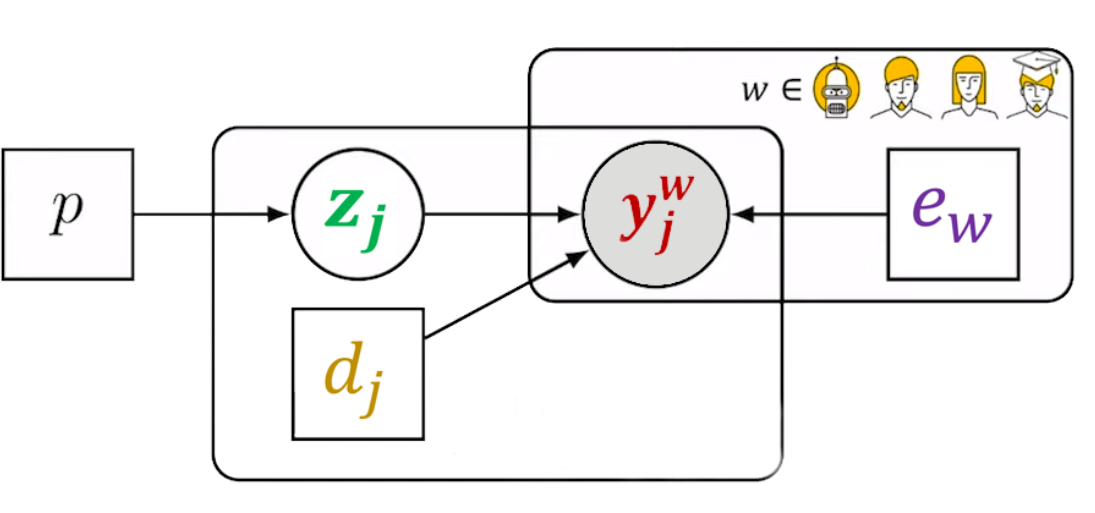


Рисунок 11 – Изображение генеративных процессов

На рисунке 11 представлено компактное изображение генеративных процессов. Круги на данных диаграммах являются случайными переменными, закрашенные круги являются наблюдаемыми переменными (шумные оценки исполнителей), белые круги – некоторая скрытая (латентная) переменная, от которой, как предполагается, зависит оценка исполнителя, квадраты – параметры генеративного процесса, а стрелки обозначают зависимости между различными параметрами системы.

Истолкование процесса, представленного на рисунке 11, следующее: предположение состоит в том, что существует некоторое условное распределение, согласно которому генерируются наблюдаемые оценки. Так, например, наблюдаемая оценка зависит от истинной оценки , которые неизвестна и, при условии этой оценки, оценка исполнителей зависит от двух параметров – сложность оценки и экспертиза исполнителя.

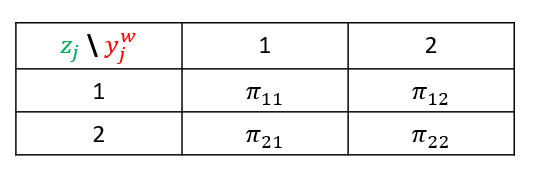


Рисунок 12 – Квадратная вероятностная матрица условного распределения

Формально, такую условную вероятность можно записать с помощью квадратной вероятностной матрицы (рисунок 12), в которой каждая строчка суммируется к единице и каждый элемент данной матрицы зависит от (введённых выше) двух параметров:

С помощью рисунка 11 восстановим генеративный процесс, согласно которому (как предполагается) были получены наблюдаемые оценки:

1. Сначала считается, что генерируется некоторая истинная оценка объекта из некоторого априорного распределения.
2. Данная оценка не наблюдается, но в зависимости от значения этой оценки приходит понимание, из какого распределения та оценка, которую поставил исполнитель. В задаче классификации это неизвестные мультиномиальные распределения.

В полученном процессе есть неизвестные параметры и для того, чтобы восстановить истинные оценки, эти параметры необходимо оценивать. Для этого делается допущение о том, что исполнители делают ошибки независимо, то есть наблюдаемые оценки при условии истинных оценок и параметров генеративного процесса являются независимыми. С помощью данного допущения записывается правдоподобие наблюдаемых и истинных оценок:

Данное правдоподобие хотелось бы максимизировать по параметрам и это стандартная задача машинного обучения, однако в данной задаче данную функцию придётся максимизировать по огромному количеству переменных, что является вычислительно сложной задачей. Поэтому при решении подобных задач используется EM-алгоритм(Expectation-maximization (EM) algorithm).

Главной идеей EM-алгоритма является следующая идея: будем максимизировать не правдоподобие выборки, а математическое ожидание логарифма правдоподобия, являющееся нижней границей оценки правдоподобия. Соответственно, будет максимизироваться и правдоподобие выборки.

EM-алгоритм состоит из двух частей:

1. **E-**шаг (Expectation шаг). На данном этапе зафиксированные параметры это априорное распределение классов, сложность заданий и экспертиза исполнителей. Если это первая итерация EM-алгоритма, то данные параметры должны быть оценены по выборке. При этих параметрах пересчитываются апостериорные распределение истинных оценок:
2. **M**-шаг (Maximization шаг). На данном шаге считается, что все истинные оценки зафиксированы в соответствие с апостериорным распределением и подбираются параметры, которые максимизируют ожидание логарифма :

Шаги 1-2 являются итерацией EM-алгоритма. У данного алгоритма есть очень важное свойство – на каждой итерации EM-алгоритма логарифм правдоподобия увеличивается, следовательно, происходит спуск на каждой итерации в сторону оптимума.

На данной идее построены два известных генеративных процесса, которые будут последовательно рассмотрены ниже.

**2.2.2. Dawid & Skene алгоритм (DS)**

Дэвид и Скин [2] рассматривают надёжность исполнителя в более широком смысле. Авторы предложили для каждого исполнителя вычислять квадратную матрицу *индивидуальных ошибок* (Individual Error Rates) размера :

Каждый элемент матрицы является вероятностью того, что исполнитель поставит метку при условии, что настоящей меткой является . Также в данной модели в качестве параметра априорного распределения классов задаётся вектор длины :

Также предполагается, что распределение оценок исполнителей и распределения классов имеют некоторые мультиномиальные распределения. Если обозначить за количество раз, которое исполнитель поставил метку примеру , то на основании данных предположений возможно записать правдоподобие выборки, которое необходимо максимизировать:

Данное правдоподобие выборки максимизируется с помощью рассмотренного EM-алгоритма, состоящего из следующих двух шагов:

1. **E-шаг:**

2. **M-шаг.** Одним из сложнейших моментов в EM-алгоритме является именно M-шаг, потому что во многих случаях требуется максимизировать функцию по параметрам с помощью градиентного спуска. К счастью, в оригинальной статье [2] авторы вывели аналитическое решение, максимизирующее математическое ожидание логарифма правдоподобия выборки:

Наличие аналитического решения для M-шага EM-алгоритма является очень важной особенностью данного алгоритма агрегации оценок исполнителей: данное свойство позволяет намного быстрее решать задачу и проводить намного большее количество итераций EM-алгоритма. На выходе, помимо получения истинных оценок, данный алгоритм позволяет получить матрицу индивидуальных ошибок для каждого из исполнителей, которая может быть использована для оценки работоспособности исполнителей. Однако данный метод не использует сложность картинок, с которыми работает исполнитель. Данное улучшение предложили создатели следующей модели.

**2.2.3. Generative Model of Labels, Abilities and Difficulties (GLAD)**

В данной модели [3] создатели предложили ввести следующую параметризацию. Пусть вероятность правильного проставления оценки исполнителем зависит от двух вещей: сложность задания и надёжность работника.

Сложность задания будет моделироваться с помощью параметра . Значение означает, что задание является очень сложным и даже самые надёжные исполнители имеют вероятность ошибиться, а значение означает, что даже самый неосведомлённый исполнитель способен правильно выполнить данное задание.

Надёжность работника моделируется параметром . При исполнитель всегда отвечает правильно, при исполнитель всегда маркирует изображения неправильно (то есть он всегда может различать два класса, но всегда неправильно). При исполнитель не может различить два класса – его метки не содержат информации об истинной метке.

Тогда вероятность того, что исполнитель даст правильный ответ на задание формируются следующим образом:

Если исполнитель имеет высокую надёжность, то и вероятность проставления верного ответа при условии, что истинной меткой будет метка будет стремиться к . При высокой сложности изображения, данная вероятность будет стремиться к .

С помощью EM-алгоритма осуществляется максимизация логарифма правдоподобия выборки.

1. **E-шаг**

2. **M-шаг**

К сожалению, M-шаг в данной реализации приходится находить с помощью градиентного спуска – аналитического решения не существует, поэтому данная особенность добавляет некоторые дополнительные проблемы в реализации.

Перейдём к демонстрации результатов работы алгоритмов и подведению некоторых итогов.

**3. СРАВНЕНИЕ АЛГОРИТМОВ АГРЕГАЦИИ ОЦЕНОК ИСПОЛНИТЕЛЕЙ**

**3.1. Сравнение алгоритмов агрегации по точности верных ответов на эталонных ответах**

На языке программирования Python в данной курсовой работе были реализованы основные методы и алгоритмы агрегации оценок, а именно:

* голос большинства (MV);
* взвешенное среднее оценок (WMA);
* взвешенный голос большинства (WMV);
* взвешенный голос успешных исполнителей (WMVT);
* байесовское голосование (BV);
* Dawid & Skene (DS);
* generative model of labels, abilities, and difficulties (GLAD).

Также были предложены и реализованы методы вычисления вероятности предоставления верной оценки исполнителем с помощью следующих методов:

* зависимость от точности работника (W);
* логистическая функция зависимости от точности работника и сложности изображения (DWS);
* индикаторная сложность изображения (IDW).

Каждый алгоритм выставлял агрегированную оценку каждому из сайтов, а затем вычислялась общая точность предсказания на эталонных ответах на заданиях. Приведём полученные результаты ниже. Для набора данных AR2 необходимо было предсказать метки для 99319 сайтов на основании результатов оценок работников. Результаты получились следующими (аббревиатуры алгоритмов совпадают с введёнными выше):

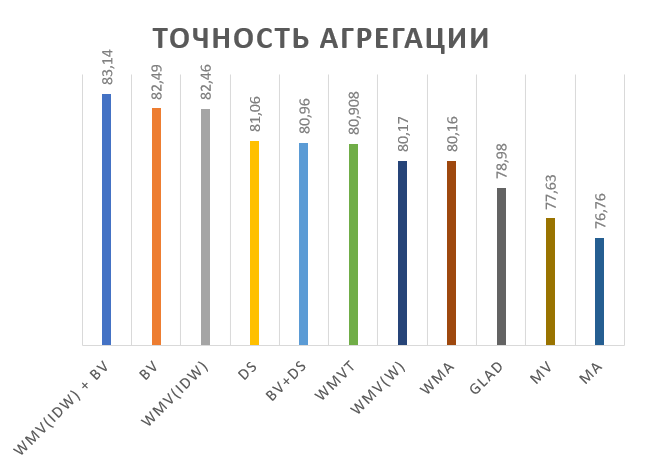


Рисунок 13 – Точность результатов агрегации реализованных алгоритмов на наборе данных AR2

Все результаты являются воспроизводимыми с той же точностью, что и на рисунке 13 за исключением алгоритма GLAD из-за разной инициализации весов сложности изображений и надёжности исполнителей. Основные выводы следующие:

* продвинутые методы агрегации дают весьма сильный прирост в качестве относительно алгоритма голоса большинства (прирост на 5-6 процентов);
* байесовское голосование является отличным улучшением результатов при известных весах (прирост на 1.5 процента);
* выбор метода проставления веса как меры надёжности исполнителя влияет на качество оценки (IDW даёт качество лучше, чем обычный способ, основанный на точности исполнителей);
* GLAD не показывает хороших результатов на наборе данных, но не стоит сбрасывать его со счетов.

Для набора данных AR5 необходимо было предсказать истинные метки для сайтов, точность оценивалась на эталонных ответах, которых для сайта могло быть несколько ввиду сложности задания.

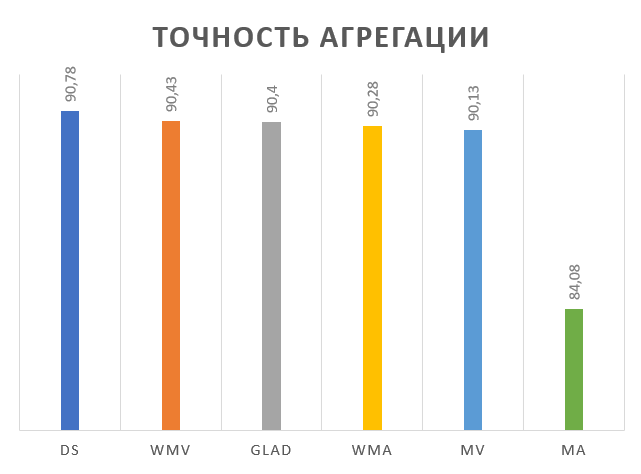


Рисунок 14 - Точность результатов агрегации реализованных алгоритмов на наборе данных AR5

На рисунке 14 видно, что модели показывают примерно один и тот же результат точности верных ответов – около 90 процентов, но всё же Dawid & Skene (DS) даёт лучший результат.

**3.2. Сравнение алгоритмов агрегации по времени работы**

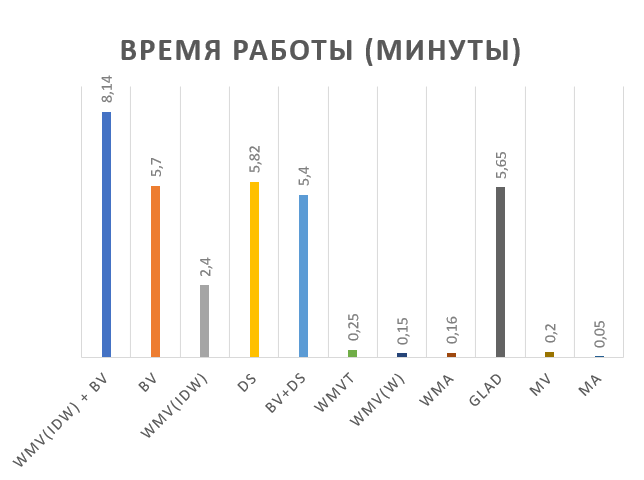


Рисунок 15 – Время работы алгоритмов агрегации алгоритмов на наборе данных AR2

На рисунке 15 видно, что алгоритмы, основанные на EM-алгоритме, являются наиболее медленными по времени работы. Время работы байесовского голосования также велико – для каждого объекта считается его правдоподобие, но, возможно, его можно улучшить с помощью более тонкой реализации. Но, как видно из рисунка 13, долгое выполнение алгоритма агрегации оправдано – все-таки точность верных ответов является более важным фактором оценки работоспособности алгоритмов агрегации.

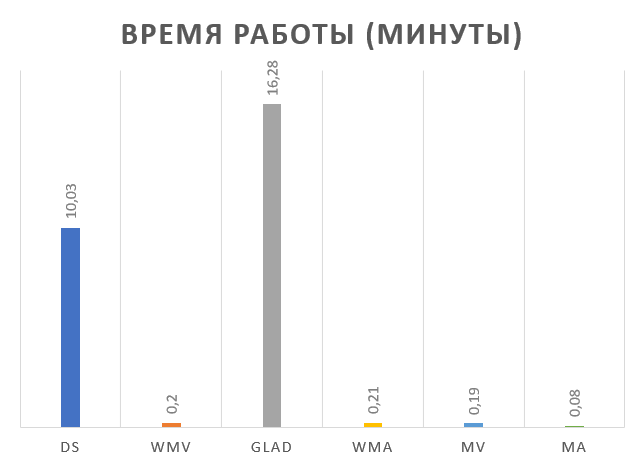


Рисунок 16 - Время работы алгоритмов агрегации алгоритмов на наборе данных AR5

На рисунке 16 видно, что два EM-алгоритма обучаются гораздо дольше остальных и гораздо дольше, чем на наборе данных AR2 из-за увеличения объёме выборки. Тем не менее, точность результатов выше, чем у голоса большинства, а значит данные алгоритмы позволяют улучшить точность агрегации оценок исполнителей.

**Заключение**

В данной курсовой работе были рассмотрены различные методы агрегации оценок исполнителей и была предложена собственная оценка надёжности исполнителей с помощью индикаторной функции сложности изображения. Все алгоритмы, описанные в данной курсовой работе, были реализованы на языке программирования Python и протестированы на работоспособность, точность и целесообразность использования – результаты точности агрегированных оценок сравнимы с заявленными результатами.

В ходе данной курсовой работы было прочитано множество статей, описывающих методы агрегации оценок исполнителей различными способами. В частности, была просмотрена модель MiniMax Conditional Entropy model (MMCE), учитывающая вероятность классов, сложность изображения и надёжность работника – все 3 компоненты, от которых зависит оценка исполнителя. В дальнейшем планируется реализация данной модели.

Также в ходе данной курсовой работы возникли трудности с реализацией моделей со скрытой оценкой. Необходимо было работать с объёмными матрицами оценок, которые необходимо было хранить в сжатом (sparse) виде, а из-за этого использовать различные приёмы для более быстрого вычисления перемножения, возведения в степень элементов sparse-матриц. В дальнейшем планируется переписывание существующего кода на более быстрый способ обработки данных.

Данная тема остаётся актуальной и современной, а значит возможны различные улучшения и доработки существующих методов.

# **Список ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Гилязев Р.А., Турдаков Д.Ю. Активное обучение и краудсорсинг: обзор методов оптимизации разметки данных. Труды ИСП РАН, том 30, вып. 2, 2018 г., стр. 215-250. DOI: 10.15514/ISPRAS-2018-30(2)-11
2. Dawid A.P., Skene A.M. Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm. Статья – Journal of the Royal Statistical Society. Series C(Applied Statistics), том 28, выпуск 1, 1979 г., стр. 20-28
3. Jacob Whitehill, Paul Ruvolo. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise: Статья – Machine Perception Laboratory, University of California, San Diego, 2009 г. – 9с.
4. Yudian Zheng, Reynold Cheng. On Optimality of Jury Selection in Crowdsourcing: Статья – Department of Computer Science, University of Hong Kong, 2015 г. – 12с.
5. Xuan Liu, Meiyu Lu, Beng Chin Ooi. CDAS: A Crowdsourcing Data Analytics System: Статья – School of Computing, National University of Singapore, 2012 г. – 12с.