# Существующие подходы к коллегиальному оцениванию. Достоинства и недостатки.

## Основные термины

МООС-платформа – сайт (платформа), на которой размещаются массовые открытые онлайн-курсы.

Обучающийся – персона, проходящая обучение на курсе.

Грейдер – обучающийся, который оценивает работы, выполненные другими обучающимися.

Рубрики – это руководство, в котором перечислены конкретные критерии оценки данного задания и уровни соответствия данному критерию с указанием начисляемых баллов за соответствие этому критерию.

Инструктор – организатор курса, который ведет лекции на курсе и проверяет задания, выполненные обучающимися.

Относительная шкала оценивания – это шкала оценивания, на основании которой оценка за работу обучающегося вычисляется относительно лучшей работы в группе.

## 1.2 Количественное коллегиальное оценивание

Количественное коллегиальное оценивание – это метод оценивания заданий, суть которого заключается в том, что оценка за выполненную работу является численным количественным значением. В настоящее время данный метод является наиболее распространенным в МООС-платформах.

Как правило, количественный подход к оценке работ используется вместе с рубриками. Рубрики очень сильно упрощают процесс оценивания, предоставляя грейдеру всю необходимую информацию для оценивания, тем самым улучшая объективность оценки, выставленной грейдером за работу. При использовании рубрик требования к навыкам оценивания у грейдера сводятся к минимуму. Так же проблемой является то, что многие студенты склонны лояльно оценивать других студентов, поэтому итоговая оценка получается менее объективной. Использование рубрик максимально унифицирует данный процесс. Нам уже не нужно полагаться на индивидуальный подход каждого студента к оцениванию работ или на навыки оценивания каждого конкретного студента, так как в рубриках можно очень подробно описать то, какими качествами должна обладать проверяемая работа для того, чтобы можно было начислить определенное количество баллов. Творческая инициатива студента при выборе количества баллов, которое он выставит за работу, сводится к минимуму.

На данный момент существенной проблемой количественного подхода к коллегиальному оцениванию является то, каким способом должна вычисляться итоговая оценка каждого обучающегося. В МООС-платформах, как Coursera, OpenSAP, в системе коллегиального PEAS, которая является расширением МООС-платформы EdX, итоговая оценка обучающегося вычисляется как среднее арифметическое по всем оценкам, которые выставили ему другие обучающиеся. Данный подход имеет довольно большую погрешность. Еще в таком способе не учитывается, что оценка обучающегося будет очень неточной, если из-за недобросовестных обучающихся, которые всем ставят чересчур завышенные оценки, он получит большую поддержку тех студентов, которые плохо оценивают. Вычисление недобросовестных студентов является первостепенной задачей алгоритмов, которые совершенствуют количественный коллегиальный подход к оцениванию работ.

Так же важной задачей не только количественного коллегиального оценивания, но и коллегиального оценивания в целом, является стимулирование студентов проверять работы качественно и добросовестно. Так как на проверку работ уходит достаточно много времени, и не у всех студентов есть желание тратить на это свои силы. Во всех МООС-платформах, которые каким-либо образом используют коллегиальное оценивание, стимулирование студентов происходит за счет начисления дополнительных баллов за проверку нескольких назначенных работ (как правило, назначают около 3 работ). Опять же при использования данного способа студент никак не контролируется. Качество обратной связи, которое он дал, может быть любым. Мы не можем знать добросовестно ли он подходил к процессу проверки работ или нет. При использовании такого подхода и добросовестным, и не добросовестным студентам начисляется одинаковое количество баллов, вне зависимости от качества проделанной ими работы. Это так же совсем не мотивирует обучающихся подходить к процессу оценивания работ ответственно. Как раз из-за отсутствия мотивации у студентов к оцениванию работ, большинство людей не рассматривает всерьез такой мощный инструмент обратной связи на МООС-платформах, как коллегиальное оценивание.

К таким методам относятся: PeerRank[ссылка], Vancouver Algorythm[ссылка]

Преимущества данного метода:

1. Возможность вводить рубрики, поясняющие тот или иной критерий оценивания. Такая возможность крайне необходима для упрощения процесса выставления баллов за задание, потому что большинство обучающихся не имеют нужных навыков выставления оценок. Так же многие студенты склонны ставить более высокие отметки своим сокурсникам. Исходя из вышесказанного, рубрики повышают объективность процесса рецензирования.
2. Инструктор получает больше информации о выполненных работах, так как использование рубрик помогает ему оценить то, какие знания и навыки студенты усвоили хуже, чтобы впоследствии на лекциях уделить этому вопросу большее внимание. В целом, данных подход увеличивает обратную связь между студентами и инструкторами. По большей части это заслуга рубрик, которые идеально подходят для их совместного использования с количественным подходом.
3. Использование абсолютной шкалы оценки заданий. Это означает, что оценки не зависят от качества работ студентов в этой группе. То есть если все студенты выполнили задание не удовлетворительно, то все они получат плохие оценки (баллы высчитываются не относительно лучшей работы в группе, как в порядковом подходе).

Недостатки данного метода:

1. Большая предвзятость у оценщиков. Для более-менее объективного оценивания нужен достаточный опыт в рецензировании работ. Каждый оценщик по-разному может оценить одну и ту же работу.
2. Абсолютная шкала оценки сама по себе довольно сложна для освоения. Людям по большей части гораздо проще сравнивать две работы и выявлять, какая из них лучше, а какая хуже (такой подход используется в порядковом методе коллегиального оценивания).

## Порядковое коллегиальное оценивание

Порядковое коллегиальное оценивание – это метод оценивания, основанный на попарном сравнении работ, который осуществляет ранжирование всех выполненных студентами заданий. Данный подход менее распространен и на практике он не применяется, однако, существует несколько статей, в которых рассматривают данный метод в теории и даже приводят формализованную математическую модель этого порядкового подхода.

Главная идея порядкового подхода к коллегиальному оцениванию заключается в том, что студент при проверке работ выставляет за работу не количество баллов, которое ему кажется подходящим, а проводит попарное сравнение нескольких пар назначенных ему работ. То есть в каждой паре он выбирает лучшую и худшую работу. Естественно, при использовании порядкового подхода использовать рубрики не представляется возможным, поэтому порядковое коллегиальное оценивание не получится применить для достаточно объемных заданий, где требуется всесторонняя проверка каждой работы.

Основным недостатком данного подхода является то, что итоговая оценка, полученная при использовании порядкового коллегиального оценивания, является не количеством баллов, а положением (рангом) в рейтинге работ всех обучающихся в группе. То есть оценка выражает не качество работы, а ее положение в рейтинге относительно лучшей работы. В случае, если все студенты в группе выполнили работу плохо, то они не получат плохую оценку.

Еще одной проблемой, которая должна решаться алгоритмами, совершенствующими порядковый подход к коллегиальному оцениванию, является то, что итоговую оценку нужно вычислить абсолютно для всех студентов, то есть необходимо охватить максимально возможное число студентов в группе. Иначе многие обучающиеся рискуют не получить оценку за задание. Это касается не только порядкового подхода, но и коллегиального оценивания в целом.

К таким методам относятся: Bayesian Peer Grading[ссылка], Mallows Model[], Bradley-Terry Model[].

Преимущества данного подхода:

1. Не требует от оценщика каких-либо специальных навыков оценивания. Для человека гораздо проще указать, что проект А хуже или лучше проекта Б, чем выставлять баллы в абсолютной шкале.
2. В некоторых математических моделях, которые использует в порядковом коллегиальном оценивании, есть специальные методы, позволяющие выявить недобросовестного оценщика.
3. Порядковый подход к оценке работ снижает нагрузку на оценщика, ему не приходится тратить дополнительное время на выставление баллов по абсолютной шкале. Следовательно, такой подход уменьшает вероятность того, что в какой-то момент ему надоест оценивать назначенные ему работы, и он выставит оценки наугад. Благодаря этому повышается качество обратной связи от оценщиков.
4. Так как люди лучше справляются со сравнением предметов, оценщик может быть в состоянии предложить более проницательные замечания по положительным и отрицательным аспектам решения студента.

Недостатки данного подхода:

1. Оценка, полученная при использовании порядкового метода, выражается в перцентилях (т.е. оценка вычисляется относительно самой лучшей работы в группе). Такая оценка несет в себе намного меньше информации о работе студента, так как оценивает не его знания, а его рейтинг в группе.
2. Необходимость переводить оценки в баллы по абсолютной шкале, потому что во всех МООС-платформах используются баллы в качестве вознаграждения за каждое выполненное задание.
3. Порядковая оценка содержит в себе меньше информации, как для инструктора, так и для оценщика. Чаще всего количественные рубрики более информативны и позволяют выяснить то, как обучающийся освоил тот или иной навык.
4. Не подходит для оценивания объемных работ таких как, эссе на более чем 150 слов и т.д.
5. Обучающимся гораздо проще оценивать, используя рубрики, т.к. они помогают прояснить некоторые детали процесса рецензирования.
6. Каждый оценщик охватывает лишь небольшую часть упорядочиваний работ.

## Сравнение количественного и порядкового подхода к коллегиальному оцениванию

В таблице 1.1 приведено сравнение количественного и порядкового подходов к коллегиальному оцениванию.

Таблица 1.1 Сравнение количественного и порядкового подходов

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Метод** | **Простота оценивания для студента** | **Методы определения недобросовестных оценщиков** | **Относительная шкала оценивания** | **Возможность использования рубрик** | **Подходит для оценивания объемных работ** | **Предполагает стимулы для грейдеров оценивать корректно** |
| **Количественное** | Требует навыков оценивания | Нет | Нет | Да | Да | Нет |
| **Порядковое** | Просто и интуитивно | Да | Да | Нет | Нет | Нет |

Исходя из этой таблицы, можно сделать вывод, что количественный метод больше подходит для большинства практических задач из-за того, что он предполагает использование рубрик, подходит для оценивания достаточно объемных работ (например, эссе на 150 слов).

Порядковый метод так же может использоваться для ряда заданий с использованием коллегиального оценивания, но из-за того, что с помощью него нельзя вычислить оценку по абсолютной шкале, и, учитывая, что в МООС-платформах обычно используется абсолютная шкала оценивания, применять этот подход будет не целесообразно.

Несмотря на то, что для коллегиального оценивания больше подходит количественный подход, в нем присутствуют несколько недостатков. Количественное оценивание никак не стимулирует грейдеров оценивать корректно, в нем отсутствуют методы наказания недобросовестных грейдеров, и так же в большинстве алгоритмов, которые высчитывают итоговую консолидированную оценку, полученную с помощью количественного подхода, используется среднее арифметическое всех оценок, которые грейдеры поставили за работу. Недостаток, связанный с вычислением итоговых баллов, как среднего от всех оценок, является самым существенным. Он сильно влияет на погрешность вычисления оценки обучающегося. Так же количественный подход меньше подходит для неподготовленных грейдеров, потому что людям гораздо проще сравнить предметы, чем выставлять оценку в баллах. Нерешенной так же является проблема слишком завышенных оценок, которые ставят студенты.

## PeerRank

PeerRank – алгоритм, который совершенствует количественный подход к коллегиальному оцениванию. Он был предложен Тоби Уолшем, профессором, изучающим искусственный интеллект, из австралийского Университета Нового Южного Уэльса[]. Название PeerRank было выбрано из-за того, что принцип работы алгоритма похож на принцип работы алгоритма ссылочного ранжирования PageRank[].

С помощью этого алгоритма оценка обучающегося вычисляется на основе оценок, которые ему поставили другие обучающиеся. При чем этот метод учитывает те оценки, которые поставил студент другим студентам, поскольку оценка – мера способности обучающегося правильно оценивать работы. Так же этот метод обеспечивает стимулы для студентов ставить оценки корректно, так как его собственная оценка зависит от оценок, которые он выставил другим обучающимся. Метод в том числе увеличивает оценку студентам, которые добросовестно подходят к процессу оценивания и уменьшает оценку тем студентам, которые недобросовестно оценивают назначенные им работы, потому что чересчур завышенные баллы, которые поставил недобросовестный грейдер могут повредить тому обучающемуся, которого оценивали, так как он получает большую поддержку тех грейдеров, которые плохо оценивают.

В большинстве систем коллегиального оценивания при вычислении итоговой консолидированной оценки студента используется среднее арифметическое от всех оценок, которые поставили ему поставили другие студенты. Большое преимущество алгоритма PeerRank в том, что с помощью него обучающиеся, которые добросовестно подходят к процессу проверки работ получают оценку выше, чем просто среднее арифметическое по всем выставленным им оценкам. Это так называемый «кредит доверия». Данный подход стимулирует обучающихся проверять работы честно. В случае, если грейдер ставит всем чересчур высокие или чересчур низкие оценки, алгоритм занижает оценку этого студента. Тоби Уолш в своей статье приводит результаты тестирования PeerRank на синтетических данных. Из них четко видно, что квадратичная погрешность алгоритма PeerRank примерно в 2 и более раз ниже, чем у метода вычисления оценок с помощью среднего арифметического.

Данный алгоритм показывает наилучшие результаты, если каждого обучающегося оценят не менее 10 других обучающихся. Скорее всего, если применять данный алгоритм на курсе, на который подписано несколько тысяч человек, то это ограничение не будет критичным. Так же для наилучшей работоспособности метода необходимо, чтобы большинство студентов добросовестно подходили к процессу проверки работ. Этот недостаток как раз предлагают решать за счет начисления «кредита доверия».

Преимущества данного алгоритма:

1. Более точное вычисление оценки, по сравнению с обычными методами вычисления консолидированной оценки, где просто вычисляется среднее между баллами, которые выставили оценщики за работу.
2. Обеспечение стимулов для студентов, для того чтобы они стремились оценивать лучше и подходили к этому процессу более ответственно. Если студент выставляет слишком заниженные баллы за работы других студентов, то его собственная оценка так же страдает от этого.
3. Использование абсолютной шкалы оценивания. Это означает, что оценка не зависит от лучшей работы в группе.
4. Данный алгоритм очень просто реализовать.
5. Метод PeerRank позволяет определить недобросовестных обучающихся, которые ставят чересчур заниженные оценки или чересчур завышенные [].

Недостатки данного алгоритма:

1. Относительно большое количество студентов, которые должны оценить каждую работу других студентов. Для нормальной работы алгоритма требуется не менее 5 оценок для каждой работы. Рекомендуемое количество рецензий на каждого студента равняется 10[]. В целом же метод не зависит от количества оценок, поставленных каждому студенту.
2. Необходимо, чтобы большинство студентов добросовестно подходило к процессу проверки работ других обучающихся.

## Bayesian Peer Grading

Bayesian Peer Grading (Байесовский метод коллегиального оценивания) – алгоритм, который совершенствует порядковое коллегиальное оценивание. Он был предложен Картником Раманом, профессором из Корнеллского Университета, который изучает машинное обучение, онлайн-обучение, Data Mining и тд.[] Данный алгоритм уже имеет готовую программную реализацию. Протестировать и посмотреть исходный код данной программы можно на веб-сайте <peergrading.com>.

Данный алгоритм использует попарные сравнения работ для упорядочивания работ по рангу. Обучающимся даются несколько пар работ других обучающихся. Они должны в каждой паре выбрать какая работа лучше. На основе этих сравнений работы упорядочиваются, а оценка вычисляется в перцентилях. То есть оценка отражает лишь положение работы в группе относительно других работ, но не отражает реальную оценку этой работы. Для каждой работы вычисляется вероятность того, какую позицию в рейтинге она будет занимать, средний ранг, медиану и энтропию.

На рисунке 1.1 изображен пример из статьи[], на котором показаны распределения вероятностей принадлежности к тому или иному рангу для 4 работ. Как видно из графиков, первая работа является лучшей из всех четырех, и из-за небольшой энтропии существует большая уверенность в том, что первая работа лучше всех остальных. У третьей работы очень большая энтропия, поэтому мы не можем с уверенностью сказать, что ее ранг равен самому высокому столбцу на графике. Для этого инструктору курса необходимо провести дополнительную проверку работы, чтобы выяснить позицию третьей работы в рейтинге более точно.

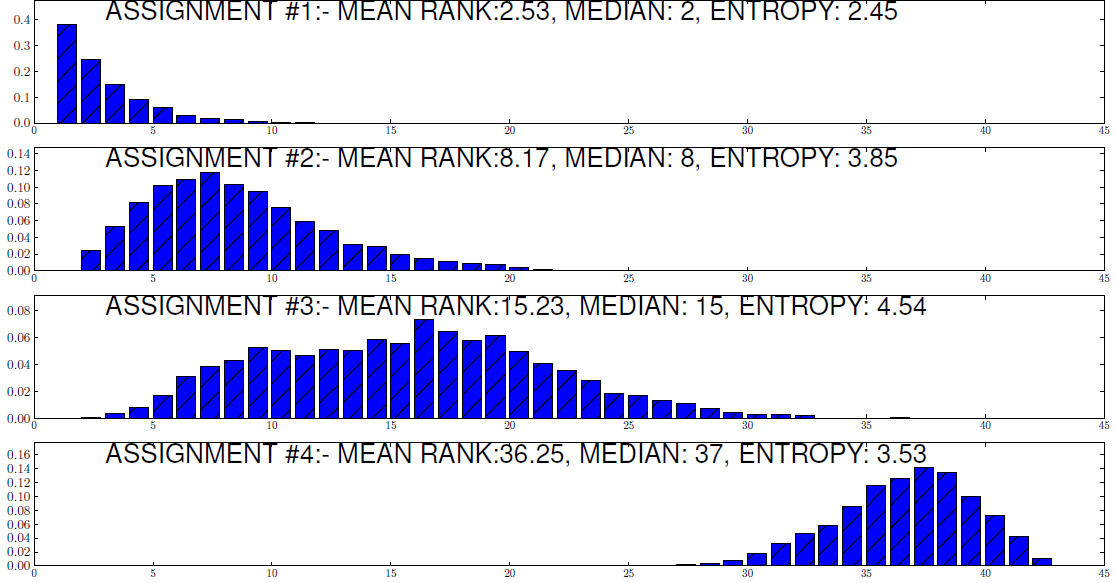


Рисунок 1.1 Пример распределения рангов для 4 работ

Данный алгоритм совершенствует уже существующие системы, которые вычисляют ранг на основе комбинации порядковой (ранговой) информации из нескольких различных источников. Аналогичный подход применяется в системах голосования («Voting Systems for Social Choice»)[], а так же применяется в системах агрегирования результатов поиска таких, как Rank Fusion и Metasearch. Такие системы используют расширения таких классических методов агрегации ранга, как модель Брэдли-Терри и модель Маллоус.

Данный алгоритм расширяет классическую модель Маллоус. В отличие от приведенных выше систем, в порядковом коллегиальном оценивании существует своя специфика агрегации рангов, имеющая фундаментальные отличия, которые, в свою очередь, не позволяют использовать уже существующие методы агрегации ранга.

Первое отличие заключается в том, что задача существующих систем агрегации поисковых результатов или систем голосования заключается в том, чтобы правильно определить ранги самых верхних элементов в рейтинге. В порядковом коллегиальном оценивании необходимо определять ранги для всей совокупности элементов (работ), а не только для лучших работ, так как итоговую оценку должны получить все обучающиеся, а не только самые успешные.

Второе отличие заключается в том, что итоговая оценка у многих обучающихся может получиться с очень высокой энтропией. Такая ситуация получается из-за того, что не все грейдеры проверяют и оценивают назначенные им работы, поэтому, наверняка, оценить работы всех обучающихся не получится или попарных сравнений с участием каких-то работ будет настолько мало, что точный ранг вычислить не представляется возможным, и из этого следует, что у вычисленного ранга данной работы будет очень высокая энтропия. Информация об энтропии так же должна быть видна инструкторам, для того чтобы они могли провести дополнительную проверку работы с целью выяснить реальное положение работы в рейтинге.

Основная сложность данного алгоритма заключается в том, чтобы по вычислить ранг работы относительно всего множества работ на основе попарных сравнений её с небольшим подмножеством работ.

Преимущества данного алгоритма:

1. Порядковый метод коллегиального оценивания более прост для понимания. Люди намного лучше справляются с попарным оцениванием работ, чем с выставлением количественной оценкой за работу, потому что для этого нужны необходимые навыки оценивания.
2. Данный алгоритм уже имеет программную реализацию и исходный код программы имеется в открытом доступе. То есть данный алгоритм существует не только в качестве теоретической математической модели.
3. Один из наиболее подробно разобранных алгоритмов, которые совершенствует порядковый подход к коллегиальному оцениванию.

Недостатки данного алгоритма:

1. Сложность данного алгоритма. Математическая модель, изложенная в статье Картника Рамана[], очень сложна для понимания человеку, не имеющего достаточной математической подготовки и знания математической статистики на очень высоком уровне. Так же автор статьи не очень много внимания уделяет тому, чтобы объяснять, назначения той или иной математической формулы, использованной в статье. В силу этой причины еще больше усиливается непонимание того, как работает данный алгоритм.
2. Для того чтобы данный алгоритм вычислял ранг достаточно точно, необходимо, чтобы для каждой работы имелось достаточно большое количество попарных сравнений данной работы с работами других обучающихся.
3. Оценка вычисляется относительно других работ в группе. Во всех МООС-платформах оценка имеет количественное выражение в баллах. В конце концов положение работы в рейтинге (ее ранг) при экспорте оценок в МООС-платформу потребуется преобразовать в итоговую оценку в баллах.