**Байесовский Порядковый Peer оценивание**

**АБСТРАКТНЫЕ**

Массивные онлайн-курсы стали доступными и

Доступный выбор для образования. Это привело к появлению новых технических

Проблемы для преподавателей, такие как оценка учащегося в масштабе.

Недавняя работа нашла порядковое сопоставление, когда отдельные

Упорядочения сортировки агрегируются в общий порядок

Из назначений, чтобы быть жизнеспособным альтернативой традиционному преподавателю /

Оценка персонала [23]. Существующие методы, которые расширяют

Методы агрегирования ранга, производят единичный заказ как выход.

Хотя эти рейтинги были признаны точными

Отражение качества задания в среднем, они не общаются

Любая неопределенность, присущая оценке

обработать. В частности, они не предоставляют преподавателям

Оценка неопределенности позиции каждого присвоения

В рейтинге. В этой работе мы решаем эту проблему, применяя

Байесовские методы к задаче градуировки по порядку,

Используя методы выборки на основе MCMC в сочетании

С моделью Мэллоуза. Эксперименты проводятся на реальном мире

Сверстников, которые демонстрируют, что предлагаемые

Обеспечивает точную информацию о неопределенности посредством

Предполагаемые задние распределения.

**Автор Ключевые слова**

Peer Оценивание, Порядковый связь, место Aggregation

**Ключевые слова ACM классификации**

H.4 Информационные системы Область применения: Разное

**Общие условия**

Алгоритмы, экспериментирование, теория

**ВВЕДЕНИЕ**

МОК (массовые открытые онлайн-курсы) предлагают обещание

Доступного высшего образования, по широте дисциплин,

Для всех, у кого есть доступ к Интернету. Вступление к

МОК вынудили инструкторов адаптировать традиционный класс

Логистику с целью охвата более 10 000 учащихся.

Одним из таких ключевых логистических аспектов является оценка

МОК. Учитывая разницу в масштабах нарядов,

Традиционные методы оценки, такие как инструктор / персонал

Для MOOC просто невозможно. В то время как существуют масштабируемые схемы автоматической классификации - например, вопросы с множественным выбором, они не подходят во всех настройках [4, 25,

12,13]. Например, для исследовательских классов требуется больше

Открытое тестирование, такое как эссе и отчеты, которые очень

Сложно оценить автоматически. Отсутствие надежной оценки

Методы для этих типов назначений могут

Виды курсов, предлагаемых как МОК.

Оценка сверстников, где студенты - не инструкторы или сотрудники -

Давать отзывы о работе других студентов в классе,

Было предложено в качестве решения. Естественное превосходство по сверстникам

Проблема масштаба [11, 16], так как число

Количество учащихся. Несмотря на это присущее

Масштабируемость классификации сверстников - ключевое препятствие для оценки сверстников

Для работы является то, что учащиеся не обучаются

Просто изучают материал сами. Для обеспечения хорошего качества

Поэтому крайне важно, чтобы рекомендации по сортировке

Легко общаться и применять, делая обратную связь

Как можно проще и проще. В направлении

Эта цель, недавняя работа предложила получить порядковый отзыв

От грейдеров [23] (например, «проект A лучше, чем проект B»)

А не кардинальные оценки (например, «проект А должен получить 87 выход

Из 100 "), поскольку порядковая обратная связь, как было показано, больше

Надежнее, чем кардинальная обратная связь [15, 3, 24, 6], и избегает

Для передачи абсолютных шкал оценки.

Это приводит к ординальной проблеме сверстников, где задано

Обратной связи грейдера (частичные упорядочения по подмножеству

Заданий), цель состоит в том, чтобы вывести общий порядок всех

Заданий. Расширены методы агрегирования ранга

К этой задаче [23] и показано, что она не только сопоставима

(Если не лучше) методов на основе кардинальной классификации

Но и традиционные методы оценки, такие как курс-персонал

(TA). Важно отметить, чем в отличие от других

Ранг агрегирования проблем, равный классификации требует точности

Во всем рейтинге, а не только на вершине.

В то время как существующие методы ординальной оценки сверстников показали, что

Оценивать рейтинги, которые в среднем точны, они просто

Вывести один рейтинг без передачи неопределенности

Присущие процессу оценки. В частности, они

Не дают преподавателям оценки неопределенности

Позиции каждого присвоения в рейтинге. Чтобы преодолеть

Это ограничение, в настоящем документе представлен метод вывода

Последующее распределение того, где каждое задание попадает в

Общий рейтинг. Эту информацию можно, например, визуализировать

Как показано на рисунке 1. Самое главное, высота

Синих полос показывает вероятность, с которой каждое присвоение

Падает на определенный ранг. Эта информация позволяет инструкторам

Убедиться в надежности алгоритма в классе

(Т.е. процентиль / позиция в рейтинге) каждого присвоения и

Различать неопределенность базовых оценок сверстников для каждого

Назначение. Например, в приведенном выше примере, хотя существует высокая вероятность того, что присвоение 1 является лучшим из четырех

Присваивания, менее вероятно, что присвоение 2 лучше, чем

3. Это связано с высокой неопределенностью в

Позиции присвоения 3 (о чем свидетельствует ее «высокая энтропия

4.54). Если такая информация предоставлена, инструкторы могут

Вмешиваться и улучшать уверенность, запрашивая дополнительные обзоры

Для конкретных заданий или, по крайней мере, с учетом

Неопределенность при выводе их оценок из рейтинга.

В этой работе мы рассматриваем проблему моделирования неопределенностей

Используя байесовские методы для классификации по порядку

Проблема. В частности, мы предлагаем Metropolis-Hastings

[8] основан метод Маркова Чейна Монте-Карло (MCMC) для

Взятие образцов с задней стороны модели Мэллоуза [20]. Результирующий

Образцы позволяют эмпирически оценить заднюю

Ранговое распределение каждого задания, что позволяет нам

Конфиденциальность и неопределенность информации.

Мы эмпирически изучаем эффективность предложенного метода на

Сверстников-сверстников, собранных из университетского уровня.

В дополнение к изучению качества ученого заднего

Упорядочения, мы также анализируем полученные в результате доверительные отношения и неопределенность

Информации, как качественно, так и количественно.

**Байесовский МЕТОДЫ ORDINAL PEER оценивания**

В этом разделе мы сначала опишем порядковое коллегиальное оценивание проблему с точки зрения машинного обучения. Затем мы кратко

обзор существующих методов для порядковое экспертной классификации проблемы. Предлагаемая байесовский версия этих методов

затем представлен, а затем эмпирической оценки этих

методы в разделе экспериментов.

**Порядковый Peer Градация (OPG) Проблема**

В порядкового задачи экспертной классификации, дано множество

| D | назначения D = {d 1, ..., d | D | } (Например, отчеты по проектам, осо-

говорит) , что нам нужно оценка. Градация выполняется

набор | G | грейдеры G = {г 1, ..., г | G | } (Например, студент-сокурсник

грейдер, рецензенты). Каждый получает грейдер подмножество назначаемых

работ D G ⊂ D для оценки. Подмножества D г может быть определена

случайным образом , путем последовательного механизма или детерминированным политика. В качестве обратной связи, каждый из грейдер обеспечивает упорядочение а (г) (возможно со связями) их назначения D г.

Основной задачей OPG является оценка порядковое оценка [ [23]](https://translate.googleusercontent.com/translate_f#8)

т.е. для получения общего упорядочения [1](https://translate.googleusercontent.com/translate_f#2) присвоений σ использование отдельных грейдер Упорядочения сг (G). В то время как мы хотели бы

Это внешкальное упорядочение σ точно соответствует некоторый (латентного) верно

приказывать о \*, мы столкнулись с несколькими проблемами. Первый,

отдельные упорядочиваний грейдер лишь частичные упорядоченности , то есть,

упорядочений охватывают лишь небольшую часть заданий

(| D г | «| D |). Вторая проблема заключается в том, что не все

грейдеры делают одинаково хорошую работу по классификации, будь то из - за усилий,

умение или понимание материала.

делать вывод, для каждого задания, процентильной ранга , как класс (общий

показатель производительности сообщает стандартизированных тестов).

Это приводит к вторичной цели надежности грейдерном оценки охватывают

ции, где мы хотели бы оценить точность / качество

η г ∈ + обратной связи каждого грейдер г. Это должно позволить

нам улучшить порядковое класс качества оценки по иден-

фикации ненадежные грейдеры и тем самым уменьшить влияние их

обратная связь по оценкам упорядочения а. Кроме того, способности для выявления ненадежных грейдеров позволяет преподавателю Incentivize хорошую и тщательную сортировку, сделав экспертную сортировку

Сама часть общего класса.

**Отношение к существующей агрегации ранга литературы**

Просмотр ординальной оценки степени в OPG можно просмотреть

Как особый вид проблемы агрегирования рангов. Агрегация рангов

[17] охватывает класс проблем, когда целью является

Комбинация порядковой (ранговой) информации из нескольких

Различных источников. Системы голосования (или «Социальный выбор» [1])

Являются одним из наиболее распространенных приложений агрегирования ранжирования

Методов. Цель этих систем - объединить предпочтения

Совокупности индивидуумов. Методы голосования Condorcet

Такие как Борда считаются среди других [10, 19], обычно

Используется для решения этих проблем. Агрегация результатов поиска

(Также известный как Ранг Fusion или Метапоиск [2]), возможно,

Наиболее известная проблема агрегирования ранга. Данные рейтинги

Из разных источников (обычно разные алгоритмы),

Цель состоит в том, чтобы объединить их и создать единый рейтинг результатов.

Расширения классических методов, таких как Маллоуз

Модель [20] и модель Брэдли-Терри [5]

Для этих проблем [18, 7] и были использованы для улучшения

Ранжирование производительности в различных настройках [22, 26, 21].

В то время как наша работа также расширяет классическую модель Маллоус,

Есть некоторые фундаментальные отличия от этих других

Проблемы агрегирования рангов, приводящие к нарушению существующих методов

Для проблемы OPG. В первую очередь это факт

Что, хотя успех агрегации и голосования результатов поиска

Системы зависят от правильной идентификации верхнего элемента (ов),

В порядковом определении оценки необходимо точно оценить

Полный рейтинг. Другими словами, мы не можем себе позволить

Любой худший из работы, определяющий 50-уе назначение перцентиля

Чем мы определяем высшие задания.

Вторая ключевая разница (и основное внимание в этой работе) - это

Тот факт, что, в отличие от других проблем агрегирования ранжирования,

Заказ заданий может быть недостаточным для этой цели

Определения сортов. Перед определением итоговых оценок

Инструкторов хотели бы иметь доступ к другим

Информации, такой как неопределенность в ранге задания.

Другими словами, они хотели бы узнать больше о распределении г (ф)

D (например, визуализация, такая как

Рисунок 1).

**Существующие подходы к OPG**

Различные подходы [[23](https://translate.googleusercontent.com/translate_f#8) ] к проблеме OPG включают расшире-

ний классических моделей , таких как Маллоус и Bradley-

Терри модель. Мы сосредоточены на методах Маллоус основе,

так как они формируют основу для методов , предложенных в этом

Работа. В частности, предлагаемый Маллоус на основе взаимного оценивания модель определяет распределение по ранжирования в терминах

Kendall-Tau расстояние [ [14]](https://translate.googleusercontent.com/translate_f#8) от истинного рейтинга о \* ассо-

signments.

D EFINITION 1. Kendall-т расстояние б K

между

ранжирование сг 1 и σ 2 это число неправильно упорядоченных пар

между двумя ранжировании и задается

δ K (σ 1, σ 2) =

Σ

d 1 ≻ σ1 d 2

Я [[d 2 ≻ σ 2 d 1]].

(1)

Учитывая грейдер Упорядочения σ (г), мы можем определить данные вероятность (если общий рейтинг был σ) , как

P ({σ (г); ∀g} | σ) =







Π

g∈G

Σ

σ ~σ (г) е -δ K (σ, σ)

Z M (| D г |)







,

(2)

где нормировочная постоянная Z M легко вычислить , как

это зависит только от ранжирования длины.

Z M (K) =

К

Π

я = 1

(

1 + е -1 + ··· + е - (я-1))

знак равно

К

Π

я = 1

1 - е -i

1 - е -1

(3)

Заметим , что в уравнении [2](https://translate.googleusercontent.com/translate_f#3) , связи в грейдер рейтинге являются модули Mod

eled как безразличие (т.е. агностиком либо рейтинга), который

приводит к суммированию в числителе берется по всем общей оцененным работам сг согласуется со слабой упорядочивания а (г). В то время как расчитывается в максимально правдоподобной оценщик (MLE) уравнения [2](https://translate.googleusercontent.com/translate_f#3)

является NP-жесткий [ [10]](https://translate.googleusercontent.com/translate_f#8) , несколько простых и податливые приближения

которые представлены хорошо работать на практике представлены в работе [ [23]](https://translate.googleusercontent.com/translate_f#8) .

Хотя эта модель не производит грейдер надежности оце-

Соседи по квартире , расширение модели предложен в [[23](https://translate.googleusercontent.com/translate_f#8) ] и COM-

рассчитан- использованием оценщик MAP (а не MLE оценки):

P ({σ (г); ∀g} | σ, {η г}) =







Π

g∈G

Σ

σ ~σ (г) е -η г δ K (σ, σ)

Z M (η г, | D г |)







,

Тем не менее, обе модели (с и без оценок надежности)

страдают от того же вопроса, в том , что они оба производят точку

По оценкам т.е. единый рейтинг в качестве выходного сигнала. В следующем разделе,

мы будем предлагать и изучать байесовский версию этих современ

ELS , который оценивает апостериорное распределение предсказанное

ранжирования и надежностей.

**Маллоус MCMC** с **использованием Метрополиса-Гастингса**

Чтобы помочь предоставить более подробную информацию для инструкторов, мы

хотели бы иметь доступ к распределению заднему

упорядоченности. Другими словами, вместо того , чтобы вероятность данных зада-

Способность мы имеем в уравнении 2 (Без учета грейдер надежностей

Алгоритм 1 Отбор проб из Маллоус Posterior с использованием

Метрополис-Гастингс

1: Входной сигнал: грейдер Упорядочения σ (г), грейдер надежностей η г и

MLE упорядочение σ.

2: Предварительно Compute х И.Я.

←

Σ

g∈G п £ I [d я

≻ σ (г)

d J] -

Σ

g∈G η г I [d J ≻ σ (г) d I [

3: σ 0 ← σ ⊳ Initialize цепь Маркова с использованием оценки MLE

4: для т = 1 ... T Do

5:

Образец σ из (Маллоус) распределения прыжков:

J MAL (σ | σ т-1)

6:

Compute отношение г т =

P (σ | {σ (г); ∀g})

P (σ т-1 | {σ (г); ∀g})

используя Уравнение [5](https://translate.googleusercontent.com/translate_f#4)

7:

С вероятностью мин (г т, 1), σ т ← σ еще σ т ← σ т-1

8:

Добавить сг т к образцам (если обкатки в условиях и прореживания

встретил)

на данный момент), мы хотели бы знать , апостериорное распределение

выведенный рейтинги σ т.е., P (σ | {σ (г); ∀g}). Мы можем с уверенностью

Предположим , равномерное предварительное на всех упорядочений (для академических fair-

Несс), что дает нам

P (σ | {σ (г); ∀g}) =

P ({σ (г); ∀g} | σ) P (σ)

Σ

σ ∈π (D) P ({σ (г); ∀g} | σ) P (σ)

знак равно

P ({σ (г); ∀g} | σ)

Σ

σ ∈π (D) P ({σ (г); ∀g} | σ)

,

(4)

С апостериорного распределения в руке, мы можем получить Умень-

хряков маргинальные распределения ранг каждого задания, или мы

может предсказать ни одного рейтинга , что сводит к минимуму задний ожидается

потеря.

Однако точные вычисления с этой задней являются infeasi-

BLE дали комбинаторное число возможных упорядочений

все назначения. Чтобы помочь нам выяснить информацию от

задний, мы будем использовать выборку на основе MCMC. Марков

Сеть Монте - Карло (или MCMC короче) представляют собой набор тех-

niques для выборки из распределения путем построения

Цепь Маркова , которая сходится к требуемому распределению

асимптотически. Метрополис-Гастингс является специфическим MCMC

алгоритм , который особенно часто встречается , когда основной

распределение трудно выборку из (как в данном случае)

особенно для распределений мульти- случайная величина.

Таким образом , чтобы помочь нам оценить свойства задней мы будем

дизайн цепи Маркова, стационарная распределение является

Распределение интереса: P (σ | {σ (г); ∀g}). Наряду с в-

ретическое гарантии , сопровождающих эти методы, добавленное

Преимуществом является то , что мы можем контролировать нужную оценить исходя

Точность ции (путем выбора количества образцов).

Это приводит к простой и эффективный алгоритм, показанный на Al-

применяемый алгоритм [1.](https://translate.googleusercontent.com/translate_f#4) Для начала мы предварительно вычислительными статистику сети

накопленная взвешенная сумма каждое задание d я не попал в статистику выше

другое назначение d J. Затем мы инициализируем цепь Маркова

используя MLE оценку упорядочения: σ. На каждом временном шаге,

предложить новый образец сг с учетом предыдущего образца σ т-1,

мы выборку из распределения прыжков (линия 5). В частности,

мы используем Маллоус на основе распределения прыжков:

Статистические данные

Плакат Отчет

Количество присвоений

42

44

Количество рецензентов 148

153

Всего экспертных обзоров

996

586

Таблица 2. Статистические данные для двух наборов данных "Плакат" и "Отчет"

,

→

J MAL (σ | σ) α е -δ K (σ, σ).

Это простое распределение для выборки из и может быть

сделано эффективно | D | войти | D | время. Кроме того , так как это

является распределение симметричным джампинг (т.е., J MAL (σ | σ) =

J MAL (σ | σ)), вычисление соотношения приема упрощается.

Когда речь идет о вычислении (акцепт) отношение R , T (Line

6), мы можем полагаться на предварительно вычисленных статистических данных сделать это эф-

тивно. В частности, мы можем упростить выражение для

Отношение к:

P (σ а | {σ (г); ∀g})

P (σ B | {σ (г); ∀g})

знак равно

Π

g∈G

абД K (σ (г), σ б) -δ K (σ (г), σ а)

знак равно

Π

I, J

е х IJ (I [d я ≻ σa d J] -I [d я ≻ σb d J])

(5)

Это выражение снова просто вычислить и может быть сделано

за время , пропорциональное количеству перевернутых пар между

σ а и σ B, что в худшем случае O (| D | 2). В целом,

Алгоритм имеет наихудший временную сложность O (T | D | 2).

Полученные образцы , полученные с помощью алгоритма может быть использован

для оценки распределения задних включая предельные

задняя ранга каждого , т.е. присваивания, Р (г d | {σ (г); ∀g},

а также статистические данные , такие как энтропия маргинальный, то

задний среднее и медиану и т.д.

Для того , чтобы улучшить качество получаемых оценок, мы

обеспечить надлежащее смешивание путем охвата умеренной скорости приема

и прореживание образцов (в наших экспериментах мы тонкие каждые 10

итераций). Кроме того , мы проводим образцы , как только цепь имеет

начали сходящихся т.е. мы используем выгорание составляет около 10000

итераций.

Мы также выведем на основе расширения Метрополиса-Гастингса из

Маллоус модель с грейдер надежностей. В дополнение к образ-

Pling упорядочений, мы также попробовать надежностей с использованием

Гауссовость прыжки (также симметрично). Однако

вычисление отношения принятие в настоящее время более активное участие и

следовательно , менее эффективно , чем для алгоритма [1](https://translate.googleusercontent.com/translate_f#4) , но тем не

меньше , может быть вычислена достаточно эффективно. Мы опускаем точное

Уравнение и вычисления для целей краткости.

Программное обеспечение и онлайн - сервис , который реализует этих техно-

ODS можно найти на http://www.peergrading.org/.

**ЭКСПЕРИМЕНТЫ**

В этом разделе мы эмпирически оценить производительность

байесовский Маллоус на основе метода экспертной классификации. В частно-

улар, мы изучаем) качество своих предсказанных рейтинга в COM-

Заготовку с существующими методами экспертной классификации , как измерено с

привет обычных классов инструктора; и б) точность

доверительных интервалов и информации о неопределенности

|  |
| --- |
| **Страница 1** |

**Байесовский Порядковый Peer Градация**

Картик комбинационного

Факультет компьютерных наук

Cornell University, Ithaca NY 14853

karthik@cs.cornell.edu

Торстен Joachims

Факультет компьютерных наук

Cornell University, Ithaca NY 14853

tj@cs.cornell.edu

**АБСТРАКТНЫЕ**

Массивные Интернет Открытые курсы стали доступны и

доступным выбором для образования. Это привело к новым техническим

вызовы для инструкторов , таких как оценка студента в масштабе.

Последние работы нашел порядковое сверстников градацию, где отдель-

UAL грейдер упорядочиваний объединяются в общую упорядочения

назначений, чтобы быть жизнеспособной заместитель традиционной инструкции

тор / оценка персонала [[23](https://translate.googleusercontent.com/translate_f#8) ]. Существующие методы, которые расширяют

методы ранга агрегации, произвести единственное упорядочение как снаружи

положил. В то время как эти рейтинги были найдены , чтобы быть точным

отражение качества назначения в среднем, они не COM-

municate любой неопределенности , присущие в оценке

обработать. В частности, они не предоставить инструкторов с

оценка неопределенности положения каждого задания по

в рейтинге. В этой работе мы будем решать эту проблему путем при-

курсирующих Байеса методы порядковому экспертной классификации зада-

методы отбора проб лем, с использованием MCMC основе в сочетании

с моделью Маллоус. Эксперименты выполнены на вещественно

мир сверстников классификации наборов данных, которые свидетельствуют о том , что про-

позировала метод обеспечивает точную информацию о неопределенности с помощью

оцененные задние распределения.

**Автор Ключевые слова**

Peer Оценивание, Порядковый связь, место Aggregation

**Ключевые слова ACM классификации**

H.4 Информационные системы Область применения: Разное

**Общие условия**

Алгоритмы, экспериментирование, теория

**ВВЕДЕНИЕ**

MOOCs (Массивные Интернет Открытые курсы) предлагают обещание

доступное высшее образование, по всей широтой дисциплин,

для тех , кто с доступом к сети Интернет. Вступление к

MOOCs заставил инструкторов адаптировать традиционную class-

номер логистики с целью масштабирования для классных комнат 10,000+ студен-

вмятины. Одним из таких ключевых логистическая является оценка студентов

MOOCs. Учитывая порядки разницы величины в масштабе,

традиционные методы оценки , такие как инструктор / штабс

на основе классификации являются просто невыгодна для MOOCs. В то время как масштабируемый

способные схемы автоматической классификации - например, множественного выбора

Разрешение на изготовление цифровых или бумажных копий всех или части этой работы для личного или

использование в классе предоставляется без гонорара при условии , что копии не сделаны или распространяться

с целью получения прибыли или коммерческой выгоды и что копии несут это уведомление и полный cita-

ции на первой странице. Авторские права на компоненты этой работы , принадлежащей другим , чем

ACM должны быть выполнены. Абстрагируясь с кредитом допускается. Для копирования в противном случае, или повторно

публиковать, размещать на серверах или перераспределять в списки, требует предварительного специального разрешения

и / или плату. Запрос разрешения от Permissions@acm.org~~pobj.

L @ S 2015, 14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада.

Авторское право принадлежит владельца / автора (ов). Права публикации лицензии на ACM.

ACM 978-1-4503-3411-2 / 15/03 ... $ 15.00.

http://dx.doi.org/10.1145/2724660.2724678

вопросы - существуют, они не подходят для всех настроек [ [4,](https://translate.googleusercontent.com/translate_f#7) 25,

[12,](https://translate.googleusercontent.com/translate_f#8) 13]. Например, проблемно-ориентированные классы требуют более

открытого состава испытаний , таких как эссе и отчетов, которые очень

сложной для автоматической оценки. Отсутствие надежной AS-

методы sessment для этих типов заданий могут ограничить

виды курсов предлагаются как MOOCs.

Peer классификации, где студенты - не преподаватели или сотрудники -

обеспечить обратную связь о работе других студентов в классе,

было предложено в качестве решения. Peer классификации естественно пере-

возникает проблема масштаба [ [11,](https://translate.googleusercontent.com/translate_f#8) 16], так как число

"Грейдеры" соответствует число студентов. Несмотря на это неиз

лор масштабируемость экспертной классификации, основным препятствием для экспертной классификации

для работы является тот факт , что студенты не обучены грейдеры и

только изучения материала сами. Для обеспечения хорошо-

оценка качества Поэтому крайне важно , чтобы классификации руководство-

линии легко общаться и применять, делая запитке

назад обработать простой и однозначна , как это возможно. В направлении

эта цель, последняя работа предложила выявляя порядковое обратной связи

от грейдеров [ [23]](https://translate.googleusercontent.com/translate_f#8) (например , "Проект А лучше проекта Б")

а не кардинальных оценок (например , "Проект А должен получить 87 из

100 "), так как порядковое обратной связи было показано, что более

надежнее , чем кардинальное обратной связи [ [15,](https://translate.googleusercontent.com/translate_f#8) 3, 24, 6], и избегает разведением

ING общаться абсолютных оценочных шкал.

Это приводит к порядковому экспертной задаче классификации, где данный

обратная связь грейдер (частичные упорядоченности над подмножеством

задания), цель состоит в том, чтобы вывести общее упорядочение всех

Назначения. Методы Ранг агрегации были экс

как правило , к этой задаче [[23](https://translate.googleusercontent.com/translate_f#8) ] и показано , что не только сравни-

BLE к методам (если не лучше , чем) Кардинал-градуировка на основе

но и традиционная практика оценки , такие как курс-персонала

(TAS) на основе классификации. Важно отметить , чем в отличие от других

проблемы агрегации ранг, равный классификации требует точности

во всем рейтинге , а не только в верхней части.

В то время как существующие методы классификации порядковое сверстников было показано

оценка рейтинга, которые точны в среднем, они просто

Выход один рейтинг без общения шается

Таинты присущие процессу оценки. В частности, они

не предоставляют инструкторов с оценкой неопределенности

позиции каждого задания в рейтинге. Для преодоления

это ограничение, этот документ представляет собой способ выведения

апостериорное распределение , где каждое задание попадает в

общий рейтинг. Эта информация может, например, быть VI-

sualized , как показано на рисунке [1.](https://translate.googleusercontent.com/translate_f#2) Самое главное, что высота

синих баров показывает вероятность , с которой каждая AS-

signment падает в определенном ранге. Эта информация позволяет не-

structors , чтобы установить доверие алгоритма в классе

(т.е. процентиль / позиция в рейтинге) каждого задания и

различать неопределенность основных классов равных для каждого

Назначение. Например, в приведенном выше примере, в то время как там

**L @ S 2015 • Градация**

**14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада**

**149**

|  |
| --- |
| **Страница 2** |

0

5

10

15

20

25

30

35

40

45

0.0

0,1

0,2

0,3

0,4

ЗАДАНИЕ # 1: - MEAN RANK: 2,53, медиана: 2, энтропией: 2,45

0

5

10

15

20

25

30

35

40

45

0.00

0,02

0.04

0.06

0,08

0.10

0,12

0,14

ЗАДАНИЕ № 2: - MEAN RANK: 8,17, медиана: 8, энтропией: 3,85

0

5

10

15

20

25

30

35

40

45

0.00

0,02

0.04

0.06

0,08

ЗАДАНИЕ № 3: - MEAN RANK: 15.23, медиана: 15, энтропией: 4,54

0

5

10

15

20

25

30

35

40

45

0.00

0,02

0.04

0.06

0,08

0.10

0,12

0,14

0,16

ЗАДАНИЕ № 4: - MEAN RANK: 36.25, медиана: 37, энтропией: 3,53

Рисунок 1. Вывод Байеса порядкового метода экспертной классификации , предложенной в данной работе. Имея алгоритм сортировки равный производить более детальный

информация о производительности каждого отдельного назначения может быть очень полезна для инструкторов , когда речь идет об определении итоговой оценки. Вышесказанное

фигура является одним из таких примеров, где для каждого задания задний маргинальное распределение (более позиции в общем рейтинге) показан (ранг на оси х,

предельная вероятность на оси у) наряду со статистическими данными , такими как задним среднее значение, медиана и энтропия маргинального распределения.

высокая вероятность того, что назначение 1 является лучшим из четырех

назначения, он менее уверен , что назначение 2 лучше , чем

Задание 3. Это происходит из - за высокой неопределенности в

Положение присвоения 3 ( о чем свидетельствует его «высокой энтропии

4,54). Если представлены с такой информацией, инструкторы могли

вмешаться и улучшить определенность путем привлечения дополнительного ре-

представления для выполнения конкретных заданий, или по крайней мере учета

неопределенность при выводе их оценки из рейтинга.

В этой работе мы рассматриваем проблему моделирования неопределенности

с использованием Байеса методы для порядкового пэра постепен-

ING проблемы. В частности , мы предлагаем Метрополиса-Гастингса

[[8](https://translate.googleusercontent.com/translate_f#8) ] метод , основанный цепь Маркова Монте-Карло (MCMC), для

отбор проб из задней модели Маллоус [ [20]](https://translate.googleusercontent.com/translate_f#8) . Там-

кающего образцы позволяют нам эмпирически оценить кзади

Оценка распределения каждого задания, что позволяет нам сообщить

откровенности и информационная неопределенность.

Мы эмпирически изучить эффективность предложенного метода на

всмотреться профилирование , наборы данных, собранных из класса на университетском уровне.

В дополнение к изучению качества ученого задней

упорядоченности, мы также анализируем полученные откровенности и не-

информация определенность, как качественно , так и количественно.

**Байесовский МЕТОДЫ ORDINAL PEER СОРТИРУЯ**

В этом разделе мы сначала опишем порядковое пэра градуировку зада-

Лем с точки зрения машинного обучения. Затем мы кратко

обзор существующих методов для порядковое экспертной классификации пробле-

Лем. Предлагаемая байесовский версия этих методов

затем представлен, а затем эмпирической оценки этих

методы в разделе экспериментов.

**Порядковый Peer Градация (OPG) Проблема**

В порядкового задачи экспертной классификации, дано множество

| D | назначения D = {d 1, ..., d | D | } (Например, отчеты по проектам, осо-

говорит) , что нам нужно оценка. Градация выполняется

набор | G | грейдеры G = {г 1, ..., г | G | } (Например, студент-сокурсник

грейдер, рецензенты). Каждый получает грейдер подмножество назначаемыми

менты D G ⊂ D для оценки. Подмножества D г может быть определена

случайным образом , путем последовательного механизма или детерминированным кажущиеся

ледяной. В качестве обратной связи, каждый из грейдер обеспечивает упорядочение а (г) (воз-

Сибли со связями) их назначения D г.

Основной задачей OPG является оценка порядковое оценка [ [23]](https://translate.googleusercontent.com/translate_f#8)

т.е. для получения общего упорядочения [1](https://translate.googleusercontent.com/translate_f#2) присвоений σ US-

ING отдельных грейдер Упорядочения сг (G). В то время как мы хотели бы

Это внешкальное упорядочение σ точно соответствует некоторый (латентного) верно

приказывать о \*, мы столкнулись с несколькими проблемами. Первый,

отдельные упорядочиваний грейдер лишь частичные упорядоченности , то есть,

упорядочений охватывают лишь небольшую часть заданий

(| D г | «| D |). Вторая проблема заключается в том, что не все

грейдеры делают одинаково хорошую работу по классификации, будь то из - за усилий,

умение или понимание материала.

1 Производство общее упорядочение заданий может быть использована для

делать вывод, для каждого задания, процентильной ранга , как класс (общий

показатель производительности сообщает стандартизированных тестов).

**L @ S 2015 • Градация**

**14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада**

**150**

|  |
| --- |
| **Page 3** |

G, G (∈ G)

Набор всех грейдеров, Удельная грейдер

D, d (∈ D)

Совокупность всех заданий, конкретное задание

D г (⊂ D)

Набор элементов , градуированных по грейдер г

σ (г)

Оценка обратной связи (с возможными связями) от г

η г (+ ∈)

Прогнозируемая надежность грейдер г

г (σ)

d

Ранг присвоения D в заказе сг (ранг 1 лучше)

d 2 ≻ σ d 1

d 2 является предпочтительным / ранг выше , чем 1 г (по а)

π (A)

Набор всех рейтингов над A ⊆ D

σ 1 ~ σ 2

∃ способ разрешения связей в а 2 для получения сг 1

σ

Расчетное упорядочение заданий

σ \*

(Латентной) Истинное упорядочение заданий

Таблица 1. Обзор Обозначения и ссылки.

Это приводит к вторичной цели надежности грейдерном оценки охватывают

ции, где мы хотели бы оценить точность / качество

η г ∈ + обратной связи каждого грейдер г. Это должно позволить

нам улучшить порядковое класс качества оценки по иден-

фикации ненадежные грейдеры и тем самым уменьшить влияние их

обратная связь по оценкам упорядочения а. Кроме того, abil-

ности для выявления ненадежных грейдеров позволяет преподавателю In-

centivize хорошую и тщательную сортировку, сделав экспертную сортировку

Сама часть общего класса.

**Отношение к существующей агрегации ранга литературы**

Порядковый задача оценивания ранга в OPG можно посмотреть

как специфический вид задачи агрегации ранга. Ранг агрес-

купности [[17](https://translate.googleusercontent.com/translate_f#8) ] охватывает класс задач , где цель

сочетание порядкового (рейтинг) информации из мульти-

PLE различных источников. Системы голосования (или социального выбора [ [1]](https://translate.googleusercontent.com/translate_f#7) )

являются одним из наиболее распространенных приложений агрегации ранга

методы. Цель этих систем заключается в объединении цен предпочтительнее

почтений из множества индивидов. Методы голосования Кондорсе

такие как Борда рассчитывать среди других [[10](https://translate.googleusercontent.com/translate_f#8) , 19], как правило ,

используется для решения этих проблем. Результат поиска Aggregation

(также известный как ранг Fusion или Метапоисковые [2 [])](https://translate.googleusercontent.com/translate_f#7) , возможно ,

наиболее хорошо известная проблема ранга агрегации. Учитывая rank-

ь из разных источников ( как правило , разные алгоритмы),

цель состоит в том, чтобы объединить их и производят один выход rank-

ING. Расширения классических методов , таких как Маллоус

модель [ [20]](https://translate.googleusercontent.com/translate_f#8) и модель Брэдли-Терри [5] стали попу-

LAR для этих проблем [[18](https://translate.googleusercontent.com/translate_f#8) , 7] и были использованы для улучшения

рейтинга производительности в различных условиях [ [22,](https://translate.googleusercontent.com/translate_f#8) 26, 21].

В то время как наша работа также расширяет классическую модель Маллоус,

Есть некоторые кардинальных отличий от этих других

проблемы агрегации ранга, которые делают существующие методы жестокого

подходит для задачи OPG. Прежде всего , это тот факт ,

что в то время как успех агрегации результатов поиска и голосующих

ING системы зависят от правильного определения верхний элемент (ы),

в оценке порядкового сорта крайне важно точно оце-

спариваться полный рейтинг. Другими словами, мы не можем позволить себе сделать

хуже без работы , идентифицирующим 50 - й процентиль назначения

чем мы выявления лучших заданий.

Вторым ключевым отличием (и основное внимание в этой работе) является

тот факт , что в отличие от других проблем агрегации ранга, един-

гле упорядочение заданий может не хватить для целей

определения классов. Прежде чем определить итоговые оценки

заданий, преподаватели хотели бы иметь доступ к другим

такая информация, как неопределенность в ранге назначаемыми

ния. Другими словами, они хотели бы знать больше о

распределение г

(Σ)

d

(например , визуализации , такие как

Рисунок [1)](https://translate.googleusercontent.com/translate_f#2) .

**Существующие подходы к OPG**

Различные подходы [[23](https://translate.googleusercontent.com/translate_f#8) ] к проблеме OPG включают расшире-

сий классических моделей , таких как Маллоус и Bradley-

Терри модель. Мы сосредоточены на методах Маллоус основе,

так как они формируют основу для методов , предложенных в этом

Работа. В частности, предлагаемый Маллоус на основе взаимного постепен-

ING модель определяет распределение по ранжирования в терминах

Kendall-Tau расстояние [ [14]](https://translate.googleusercontent.com/translate_f#8) от истинного рейтинга о \* ассо-

signments.

D EFINITION 1. Kendall-т расстояние б K

между

ранжирование сг 1 и σ 2 это число неправильно упорядоченных пар

между двумя ранжировании и задается

δ K (σ 1, σ 2) =

Σ

d 1 ≻ σ1 d 2

Я [[d 2 ≻ σ 2 d 1]].

(1)

Учитывая грейдер Упорядочения σ (г), мы можем определить данные вероят

капот (если общий рейтинг был σ) , как

P ({σ (г); ∀g} | σ) =







Π

g∈G

Σ

σ ~σ (г) е -δ K (σ, σ)

Z M (| D г |)







,

(2)

где нормировочная постоянная Z M легко вычислить , как

это зависит только от ранжирования длины.

Z M (K) =

К

Π

я = 1

(

1 + е -1 + ··· + е - (я-1))

знак равно

К

Π

я = 1

1 - е -i

1 - е -1

(3)

Заметим , что в уравнении [2](https://translate.googleusercontent.com/translate_f#3) , связи в грейдер рейтинге являются модули Mod

eled как безразличие (т.е. агностиком либо рейтинга), который

приводит к суммированию в числителе берется по всем общей заказа-

ь сг согласуется со слабой упорядочивания а (г). В то время как расчетными

ING в максимально правдоподобной оценщик (MLE) уравнения [2](https://translate.googleusercontent.com/translate_f#3)

является NP-жесткий [ [10]](https://translate.googleusercontent.com/translate_f#8) , несколько простых и податливые приближения

которые представлены хорошо работать на практике представлены в работе [ [23]](https://translate.googleusercontent.com/translate_f#8) .

Хотя эта модель не производит грейдер надежности оце-

Соседи по квартире , расширение модели предложен в [[23](https://translate.googleusercontent.com/translate_f#8) ] и COM-

рассчитан- использованием оценщик MAP (а не MLE оценки):

P ({σ (г); ∀g} | σ, {η г}) =







Π

g∈G

Σ

σ ~σ (г) е -η г δ K (σ, σ)

Z M (η г, | D г |)







,

Тем не менее, обе модели (с и без оценок надежности)

страдают от того же вопроса, в том , что они оба производят точку

По оценкам т.е. единый рейтинг в качестве выходного сигнала. В следующем разделе,

мы будем предлагать и изучать байесовский версию этих современ

ELS , который оценивает апостериорное распределение предсказанное

ранжирования и надежностей.

**Маллоус MCMC** с **использованием Метрополиса-Гастингса**

Чтобы помочь предоставить более подробную информацию для инструкторов, мы

хотели бы иметь доступ к распределению заднему

упорядоченности. Другими словами, вместо того , чтобы вероятность данных зада-

Способность мы имеем в уравнении 2 (Без учета грейдер надежностей

**L @ S 2015 • Градация**

**14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада**

**151**

|  |
| --- |
| **Страница 4** |

Алгоритм 1 Отбор проб из Маллоус Posterior с использованием

Метрополис-Гастингс

1: Входной сигнал: грейдер Упорядочения σ (г), грейдер надежностей η г и

MLE упорядочение σ.

2: Предварительно Compute х И.Я.

←

Σ

g∈G п £ I [d я

≻ σ (г)

d J] -

Σ

g∈G η г I [d J ≻ σ (г) d I [

3: σ 0 ← σ ⊳ Initialize цепь Маркова с использованием оценки MLE

4: для т = 1 ... T Do

5:

Образец σ из (Маллоус) распределения прыжков:

J MAL (σ | σ т-1)

6:

Compute отношение г т =

P (σ | {σ (г); ∀g})

P (σ т-1 | {σ (г); ∀g})

используя Уравнение [5](https://translate.googleusercontent.com/translate_f#4)

7:

С вероятностью мин (г т, 1), σ т ← σ еще σ т ← σ т-1

8:

Добавить сг т к образцам (если обкатки в условиях и прореживания

встретил)

на данный момент), мы хотели бы знать , апостериорное распределение

выведенный рейтинги σ т.е., P (σ | {σ (г); ∀g}). Мы можем с уверенностью

Предположим , равномерное предварительное на всех упорядочений (для академических fair-

Несс), что дает нам

P (σ | {σ (г); ∀g}) =

P ({σ (г); ∀g} | σ) P (σ)

Σ

σ ∈π (D) P ({σ (г); ∀g} | σ) P (σ)

знак равно

P ({σ (г); ∀g} | σ)

Σ

σ ∈π (D) P ({σ (г); ∀g} | σ)

,

(4)

С апостериорного распределения в руке, мы можем получить Умень-

хряков маргинальные распределения ранг каждого задания, или мы

может предсказать ни одного рейтинга , что сводит к минимуму задний ожидается

потеря.

Однако точные вычисления с этой задней являются infeasi-

BLE дали комбинаторное число возможных упорядочений

все назначения. Чтобы помочь нам выяснить информацию от

задний, мы будем использовать выборку на основе MCMC. Марков

Сеть Монте - Карло (или MCMC короче) представляют собой набор тех-

niques для выборки из распределения путем построения

Цепь Маркова , которая сходится к требуемому распределению

асимптотически. Метрополис-Гастингс является специфическим MCMC

алгоритм , который особенно часто встречается , когда основной

распределение трудно выборку из (как в данном случае)

особенно для распределений мульти- случайная величина.

Таким образом , чтобы помочь нам оценить свойства задней мы будем

дизайн цепи Маркова, стационарная распределение является

Распределение интереса: P (σ | {σ (г); ∀g}). Наряду с в-

ретическое гарантии , сопровождающих эти методы, добавленное

Преимуществом является то , что мы можем контролировать нужную оценить исходя

Точность ции (путем выбора количества образцов).

Это приводит к простой и эффективный алгоритм, показанный на Al-

применяемый алгоритм [1.](https://translate.googleusercontent.com/translate_f#4) Для начала мы предварительно вычислительными статистику сети

накопленная взвешенная сумма каждое задание d я не попал в статистику выше

другое назначение d J. Затем мы инициализируем цепь Маркова

используя MLE оценку упорядочения: σ. На каждом временном шаге,

предложить новый образец сг с учетом предыдущего образца σ т-1,

мы выборку из распределения прыжков (линия 5). В частности,

мы используем Маллоус на основе распределения прыжков:

Статистические данные

Плакат Отчет

Количество присвоений

42

44

Количество рецензентов 148

153

Всего экспертных обзоров

996

586

Таблица 2. Статистические данные для двух наборов данных "Плакат" и "Отчет"

,

→

J MAL (σ | σ) α е -δ K (σ, σ).

Это простое распределение для выборки из и может быть

сделано эффективно | D | войти | D | время. Кроме того , так как это

является распределение симметричным джампинг (т.е., J MAL (σ | σ) =

J MAL (σ | σ)), вычисление соотношения приема упрощается.

Когда речь идет о вычислении (акцепт) отношение R , T (Line

6), мы можем полагаться на предварительно вычисленных статистических данных сделать это эф-

тивно. В частности, мы можем упростить выражение для

Отношение к:

P (σ а | {σ (г); ∀g})

P (σ B | {σ (г); ∀g})

знак равно

Π

g∈G

абД K (σ (г), σ б) -δ K (σ (г), σ а)

знак равно

Π

I, J

е х IJ (I [d я ≻ σa d J] -I [d я ≻ σb d J])

(5)

Это выражение снова просто вычислить и может быть сделано

за время , пропорциональное количеству перевернутых пар между

σ а и σ B, что в худшем случае O (| D | 2). В целом,

Алгоритм имеет наихудший временную сложность O (T | D | 2).

Полученные образцы , полученные с помощью алгоритма может быть использован

для оценки распределения задних включая предельные

задняя ранга каждого , т.е. присваивания, Р (г d | {σ (г); ∀g},

а также статистические данные , такие как энтропия маргинальный, то

задний среднее и медиану и т.д.

Для того , чтобы улучшить качество получаемых оценок, мы

обеспечить надлежащее смешивание путем охвата умеренной скорости приема

и прореживание образцов (в наших экспериментах мы тонкие каждые 10

итераций). Кроме того , мы проводим образцы , как только цепь имеет

начали сходящихся т.е. мы используем выгорание составляет около 10000

итераций.

Мы также выведем на основе расширения Метрополиса-Гастингса из

Маллоус модель с грейдер надежностей. В дополнение к образ-

Pling упорядочений, мы также попробовать надежностей с использованием

Гауссовость прыжки (также симметрично). Однако

вычисление отношения принятие в настоящее время более активное участие и

следовательно , менее эффективно , чем для алгоритма [1](https://translate.googleusercontent.com/translate_f#4) , но тем не

меньше , может быть вычислена достаточно эффективно. Мы опускаем точное

Уравнение и вычисления для целей краткости.

Программное обеспечение и онлайн - сервис , который реализует этих техно-

ODS можно найти на http://www.peergrading.org/.

**ЭКСПЕРИМЕНТЫ**

В этом разделе мы эмпирически оценить производительность

байесовский Маллоус на основе метода экспертной классификации. В частно-

улар, мы изучаем) качество своих предсказанных рейтинга в COM-

Заготовку с существующими методами экспертной классификации , как измерено с

привет обычных классов инструктора; и б) точность

доверительных интервалов и информации о неопределенности.

**L @ S 2015 • Градация**

**14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада**

**152**

|  |
| --- |
| **Page 5** |

**Экспериментальная настройка**

Мы использовали одноранговой сортируя наборов данных , введенных в работе [[23](https://translate.googleusercontent.com/translate_f#8) ]. Эти

наборы данных были собраны в условиях реального классе от а

большой университетский класс. Класс , который состоял из 170

студенты и 9 ассистентов (TAS), используемый сверстников классификации

оценить курсовых проектов (сделано в группах по 3-4 сту-

вмятины) Преимущество такого размера класса является наличие

обычные классы , основанные инструктора для выполнения заданий, в Ad-

условие для классов одноранговых (выполняется индивидуально каждым студен-

вмятина). Имея эти оценки инструктора позволяет нам предоставлять

более надежной оценки образовательного влияния этих

методы, сверх того , что предыдущая работа сделала.

Мы использовали как плакат и заключительный наборы данных отчетов в этом

Работа. Эти два набора данных соответствуют различным частям

курс. Студенты были стимулы , чтобы сделать хорошую аттестацию рабочих мест,

путем включения их сверстников производительность классификации в их

Общая оценка за курс. Сверстников классификации было сделано

10-балльной (кардинал) Лайкерта масштаб так, чтобы сравнивать кардинальные

и порядковые методы экспертной классификации. Порядковый равный постепен-

методы лишь ИНГ использовал упорядочение подразумевает кардиналом

оценки.

В таблице [2](https://translate.googleusercontent.com/translate_f#4) приводятся некоторые из ключевых статистических двух наборов данных.

В среднем каждый плакат и окончательный отчет получил примерно 24

и 13 экспертных обзоров соответственно. Для обоих наборов данных там

был один класс инструктора для каждого задания. Как де-

описано в работе [[23]](https://translate.googleusercontent.com/translate_f#8) , преподаватель классов для докладов водством

ределяется совершенно не зависит от оценок сверстников. Для

плакаты сорта инструктор использовал классы ТП, которые

были частично под влиянием студенческих классов.

Метод байесовской Маллоус MCMC был запущен с тождественным

(фиксированные) параметры кал для обоих наборов данных. В общей сложности 5000 образца

упорядочиваний были взяты из цепи Маркова с использованием алго-

горитм [1](https://translate.googleusercontent.com/translate_f#4) . Эти образцы были использованы для оценки кзади

распределения и для получения статистических данных в следующем

подразделы.

**Являются ли предполагаемые упорядочиваний точны?**

Ключевым преимуществом байесовского подхода заключается в том, что задняя

Распределение упорядочениями обеспечивает неопределенность информации

ции. Но мы также можем использовать апостериорное распределение предсказать,

единый порядок назначений. Как точность

упорядочений предсказанные байесовской модели сравнить с

Точность упорядочениями оценивается с помощью максимального правдоподобия

оценка (MLE)? Для того, чтобы ответить на этот вопрос, мы сравним

следующие методы:

• MLE: максимального правдоподобия оценщик из Маллоус

модель [23 [].](https://translate.googleusercontent.com/translate_f#8) Это оценка в одной точке, а также

используется для инициализации цепи Маркова.

• Режим-MAL: (Один из) моды заднему

Распределение Маллоус. Связи разбиты случайным образом .

• Режим-MAL + G: (Один из) моды заднему

Распределение Маллоус с оценками надежности грейдер. связи

разбиты случайным образом .

21,8

21,4

22,7

27,0

18,9

25,3

22,2

21,4

19,7

26,3

0

5

10

15

20

25

АФИША

ОТЧЕТ

MLE

Режим-

MAL

Bayes-

MAL

Режим-

MAL + G

Bayes-

MAL + G

14,9

9.4

14.7

11.5

14.3

10.8

15,3

9.4

14.3

11.2

0

4

8

12

16

АФИША

ОТЧЕТ

MLE

Режим-

MAL

Bayes-

MAL

Режим-

MAL + G

Bayes-

MAL + G

Рисунок 2. производительность (Top) Нормированный Kendall-тау всех методов

против инструктора классов для обоих наборов данных: Плакат (слева) и отчет

(Правильно). Рисунок на дне похож , но сообщает взвешенную версию

ошибка Kendall-тау. Примечание: Производительность случайной базовой линии будет

50% для обоих показателей. Для обеих фигур, тем меньше значение , тем лучше.

• Байеса-MAL: Это оценка Байеса минимизации дистальных

RIOR ожидается , K & delta над задним выученное Alg [1](https://translate.googleusercontent.com/translate_f#4) . Для-

Малли, предсказанное упорядочение

Σ = σ argmin

Σ

σ

δ K (σ, σ) P (σ | D),

где P (σ | D) представляет собой расчетную заднюю распреде-

ции ( в качестве выходного сигнала по методу байесовский MCMC).

• Байеса-MAL + G: Это оценка минимизирующий Байеса

задний ожидается Т КТ по заднему Маллоус

модель с оценками надежности грейдер.

В то время как вычисления Байеса-MAL и Байеса-MAL + G пре-

dictions является NP-трудной задачей, так как она требует вычисляя

Кемени-оптимальная совокупная [ [10]](https://translate.googleusercontent.com/translate_f#8) , мы можем приблизить онтиче-

timal решение задачи минимизации эффективно. В частно-

ности, мы использовали простой и эффективный Борда-Count тех-

Nique, которая , как известно, в 5-приближение [ [9]](https://translate.googleusercontent.com/translate_f#8) . В нашем

случай, это также несет в себе хорошую смысловую нагрузку , поскольку это составляет

просто приказывать задания их задней означают

занимает.

Результаты показаны на рисунке [2](https://translate.googleusercontent.com/translate_f#5) . В качестве меры пре-

Точность дикция, мы используем ошибку Kendall-тау в отношении

инструктора рейтинга. Мы также вычислить взвешенное Ver-

Sion ошибки Kendall-тау, где misordering элементы с

**L @ S 2015 • Градация**

**14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада**

**153**

|  |
| --- |
| **Page 6** |

44,3

74,5

17,9

31,1

0

10

20

30

40

50

60

70

80

50% Интервал

80% Интервал

**АФИША**

OVERLAP

ШИРИНА

51,7

86,0

24,4

43,0

0

10

20

30

40

50

60

70

80

50% Интервал

80% Интервал

**ОТЧЕТ**

OVERLAP

ШИРИНА

Рисунок 3. Средняя Перекрытие (сплошные зеленые полоски) от 50% и 80% Байесовских надежных интервалов с распределением инструктор ранга, для интервалов

полученный методом байесовского Маллоус MCMC для плакате (слева) и ОТЧЕТ (справа) наборов данных. Наряду с перекрытием средний размер

(ширина) интервала (в процентах) от общего рейтинга длины в (красные полосатые столбики).

больший (инструктор) разница оценка приводит к худшим производи-

Mance мера. Следует отметить , что обе эти меры нор

мированных лежать в диапазоне от 0 ( что указывает на совершенное согласие с

инструкторов) и 100 ( с указанием полного разворота из

инструктор рейтинга). На обоих наборах данных, производительность

предложенные байесовские методы не являются , по существу , различие

лор от ОМП. Там , как представляется , нет четкой тенденции

что один метод превосходит другие, и различия

, вероятно , из - за того , что сорта инструктор используется в качестве

золотой стандарт сами подвержены неопределенности. Один-

подать в суд , следует отметить, что методы "Mode" , как правило, больше ,

дисперсия, как производительность может меняться в зависимости от режима , который был

выбран (как распределение имеет тенденцию быть мультимодальным).

И, наконец, мы также отмечаем , что производительность не сильно отличается

с добавлением оценки надежности грейдер. Это наблюдение

согласен с подобным выводом , сделанные в [ [23]](https://translate.googleusercontent.com/translate_f#8) (для обоих порядкового

и методы кардинальное классификации). Наиболее вероятной причиной

наблюдая такое поведение является явным стимулом в плане

класс кредита , который студенты были даны для выполнения тщательного

Работа с коллегиальных обзоров, таким образом, что количество действительно суб-

стандартные обзоры в данных может быть низким.

**Насколько хорошо оцененные доверительные интервалы?**

В то время как предыдущий эксперимент показал , что в целом качест-

во порядками имеет тенденцию быть довольно хорошим (в отношении

для инструктора классов), он не говорит нам , насколько точно

Байесовской модели подход неопределенность предсказанное

занимает. Для того, чтобы ответить на этот вопрос, мы теперь оценить , насколько хорошо

байесовские доверительные интервалы (например, заслуживающие доверия интервалов)

выведенный задние маргинальные распределения (более позиции в

общий рейтинг) для индивидуальных заданий являются. Для оцени-

ели эти оценки неопределенности, мы снова используем инструктор

классов [2](https://translate.googleusercontent.com/translate_f#6) , В частности, мы оцениваем качество 50% и

80% достоверные интервалы.

2 Так как они также имеют связи, мы рассматриваем связи как безразличие и , следовательно ,

имеют равномерное распределение вероятностей по всем возможным действительным ранга

позиции.

Для каждого задания, мы сначала вычислить (заднюю)

маргинальный распределение по позициям рейтинга , как показано на

Рисунок [1](https://translate.googleusercontent.com/translate_f#2) из введения. Затем мы вычисляем перекрытие

из достоверных интервалов этих маргиналов с инструктором

Распределение рейтинга т.е. присвоение которого доверие интер-

вал содержит (все) инструктора при условии звания имеет 100%

перекрывают друг друга, в то время как интервал с не перекрывают друг друга забивает 0%. Мы

сообщить об этом перекрытие , усредненное по всем присвоений. К тому же

к этому, мы также сообщить размер этих интервалов (как центных

возраст общей длины рейтинга).

Результаты показаны на рисунке [3.](https://translate.googleusercontent.com/translate_f#6) Мы считаем , что интервалы

производится на основе байесовского MCMC техники Маллоус

которые хорошо откалиброваны. В частности, для обоих плакатов

и доклады, 50% и 80% интервал крышки примерно что

процент классов инструктора по желанию (как указано

значениями перекрываться). Наблюдаемое перекрытие намного больше

чем размер интервала, что указывает на прогностическую Персонов

что производительности таких гораздо лучше , чем случайное. Эти результаты показывают ,

что расчетные интервалы имеют смысл и передать точно-

Скорость передачи информации о неопределенности. Результаты при включении

Информация о надежности грейдер подобны и , следовательно , оставленный к

Во избежание дублирования.

**Как остроконечный являются распределения задних?**

Результаты на рисунке [3](https://translate.googleusercontent.com/translate_f#6) , показывают , что доверительный интервал

отчеты имеют большую ширину , чем для плакатов т.е.

существует большая неопределенность в маргинальные докладов , чем

плакаты. Это говорит о том, что распределение задних являются

более достиг максимума вокруг режима для плакатов по сравнению с

отчеты. Чтобы убедиться в этом, мы вычислили ожидаемые значения

ошибки Kendall-Tau (и взвешенной ошибки Kendall-тау)

под раздачу задней:

Σ

σ

δ K (σ \*, σ) P (σ | D)

Заметим , что а \* относится к инструктору ранжирования и Р (а | О)

ученый задний. Мы называем эти ценности , как EXP-MAL

**L @ S 2015 • Градация**

**14-18 марта, 2015, Ванкувер, Британская Колумбия, Канада**

**154**

|  |
| --- |
| **Page 7** |

18,9

25,3

20,7

20,9

19,7

26,3

21,4

30,2

0

5

10

15

20

25

30

АФИША

ОТЧЕТ

Bayes-

MAL

exp-

MAL

Bayes-

MAL + G

exp-

MAL + G

14.3

10.8

15,6

29,3

14.3

11.2

16.4

29,1

0

5

10

15

20

25

30

АФИША

ОТЧЕТ

Bayes-

MAL

exp-

MAL

Bayes-

MAL + G

exp-

MAL + G

Рисунок 4. Kendall-тау (Top) и Weighted Kendall-тау (Bottom) Пер-

влиянии на состояние байесовской оценки точки рейтинга по сравнению с ожидаемым

эффективность распределения задней рейтинга. Для обоих рисунках

более низкое значение , тем лучше.

(без грейдер надежностей) и EXP-MAL + G (с грейдера

оценка надежности). Результаты показаны на рисунке [4.](https://translate.googleusercontent.com/translate_f#7)

Мы считаем , что разница в производительности между Байеса

оценка (Байеса) и ожидаемое значение (EXP) полной

задний , как правило , больше для отчетов , чем для плакатов.

Для плакатов, кажется , что задняя настолько узок , что

почти любой образец из задней близко к Байеса

оценить. Для отчетов, задняя менее козырьком. Один

Объяснение это большее число обзоров , доступных для

плакаты.

Наконец, мы хотели бы исследовать , какие назначенные им

Метод Байеса сверстников классификации является наиболее уверены, и

как эта неопределенность относится к оценкам , данных входов

structors. Для того, чтобы обеспечить некоторое представление о том , мы вычисляем заднюю

маргинальные энтропий всех заданий, а затем усреднить

Энтропия для всех заданий с тем же самым кардинальным инструктором

класс. Результат визуализируется на рис [.](https://translate.googleusercontent.com/translate_f#7)5. Задания , которые

получают самые высокие и самые низкие баллы инструктора, как правило,

присвоения с самой низкой задней предельной энтропии.

Назначение в середине, как правило, имеют более высокие энтропию,

указывая, что метод менее уверены в своем положении

в рейтинге. Основываясь на этих выводах, наша гипотеза, что

это "легко" для студентов и преподавателей, чтобы определить

очень хорошие и очень плохие заданий. Присвоений в

средний гораздо труднее сорта, так как они требуют тщательного

2.8

3.3

3,8

4.3

4.8

4

5

6

7

8

9

10

**verage En**

**Trop**

**Y**

**Инструктор Кардинал Score**

плакат

доклад

Рисунок 5. Распределение средних энтропий маргиналов, когда

агрегируется оценок Кардинал инструктор. Очки без каких-либо погрешностями

указать только одно назначение с этим счетом.

Компромисс между различными типами ошибок. Она также может быть

так, что некоторые из назначений в середине трудно

сравнить, так как они настолько различны в теме, что значимое

сравнение затруднено.

**ВЫВОДЫ И БУДУЩАЯ РАБОТА**

В этой работе мы предложили использовать байесовские методы для

Проблема ординальной оценки сверстников, чтобы обеспечить инструкторов

Более богатая информация, сообщающая неопределенность дополнительно

К предсказанным порядковым классам. Предлагаемый нами метод

Использует пробник MCMC на базе Metropolis-Hastings для

Модель сверстников Маллов. Мы эмпирически подтвердили

Предлагаемые методы и показать выведенным потомкам согласие

С классами инструкторов и для передачи точного количества

Неопределенность.

В дополнение к дальнейшим эмпирическим исследованиям качества

Мы изучаем, как разрабатывать байесовские

Методы вывода также для других порядковых моделей выравнивания сверстников.

Открытый вопрос касается качества достоверных интервалов

Расчетной надежности грейдеров. Кроме того, существуют

Много интересных вопросов о том, как получить обратную связь

От студентов. Например, это может быть когнитивно

Менее требовательна к ученикам, чтобы нарушить их порядковую оценку

В парные сравнения [15], особенно если

Количество предметов для оценки велико.