|  |
| --- |
| **Страница 1** |

**Методы Порядковый Peer Grading**

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 имеют потенциал революциях

высшего образования с их широкой разъяснительной работы и принадлеж-

ность, но они требуют инструкторов , чтобы придумать масштабируемым

чередуется к традиционной оценке студента. Peer градация -

студенты , имеющие оценки друг друга - это перспективный подход

к решению проблемы оценки в масштабе, так как NUM-

личество "классников" , естественно , весы с числом студентов.

Тем не менее, студенты не обучены в классификации, что означает ,

что нельзя ожидать такой же уровень навыков классификации , как и в

традиционные настройки. Опираясь на широкую доказательства того, что порядковое

обратная связь проще обеспечить и надежнее , чем кардинал

Сигнал обратной связи [5, 38, 29, 9], поэтому желательно , чтобы Peer

грейдеры сделать порядковые заявления (например , "Проект Х Пари

тер , чем проекта Y ") , а не требовать от них , чтобы сделать кардинальные

заявления (например , "Проект X является B-"). Таким образом, в этой статье мы

изучить проблему автоматически умозаключения оценки студентов

от порядкового обратной связи со стороны сверстников, в отличие от существующих методов

которые требуют кардинального обратной связи со стороны сверстников. Сформулируем порядковую проблема экспертной классификации как тип агрегации ранга проблема, а также изучить несколько вероятностных моделей в соответствии с которым в

оценить оценки студентов и надежности грейдер. Изучается

применимость этих методов с использованием данных равный сортируя кол-

бранной из реального класса - с инструктором и ТА классов в качестве

базовый уровень - и демонстрируют эффективность порядковым обратной связи

методы в сравнении с существующим кардинальное сверстников классификации

методы. И, наконец, мы сравним эти пэром классификации методов

к традиционным методам оценки.

Категории и Предметные Дескрипторы

H.4 [Информационные системы] Приложения: Разное

Общие условия

Алгоритмы, экспериментирование, теория

Ключевые слова

Peer Оценивание, Порядковый связь, место Aggregation

1. ВВЕДЕНИЕ

Появление MOOCs (Массивные Интернет Открытые курсы)

обещает беспрецедентный доступ к образованию с учетом их относительно низкие затраты и широкие охват, расширение прав и возможностей обучения через

разнообразный ассортимент предметов для тех , кто с доступом к Ин-

тернете. Классы часто имеют свыше 20000 студентов,

который на несколько порядков больше , чем обычный университетский класс. Таким образом, преподаватели вынуждены пересмотреть class-

комната логистики и практики с тем, чтобы масштабировать до MOOCs.

Одним из ключевых открытых задач является оценка студента для

такие большие классы. Традиционные методы оценки, такие

в качестве инструкторов или ассистентов (TA) сортируя отдель-

UAL студенческие задания, просто неосуществимы в таком масштабе.

Следовательно, задания в большинстве современных MOOCs взять

форма простых с несколькими вариантами ответов и другие схемы

которые могут классифицироваться автоматически. Тем не менее, опираясь на такие

жесткие схемы тестирования не обязательно служить хорошим входов

dicator обучения и отстает от обычного тест-дизайна

стандарты [16, 17]. Кроме того, такое ограничительное тест-

ING методологии ограничивает результаты обучения , которые могут быть

испытания, или даже ограничивает виды курсов , которые могут быть предложены.

Например, либерал-искусства курсы и проблемно-ориентированные классы

требуют более открытые задания и ответы (например,

эссе, проектные предложения, отчеты по проектам).

Peer оценивание имеет потенциал , чтобы преодолеть ограничения

описанные выше в то время как масштабирование по размеру даже самый большой

MOOCs. В экспертной классификации, студенты - не преподаватели или

TAs - обеспечить обратную связь о работе других студентов в

их класс [14, 22], а это означает , что число " х классов"

естественно , растет с количеством студентов. В то время

масштабирования свойства экспертной классификации являются привлекательными, есть

ряд проблем в принятии экспертной классификации работы.

Одна из ключевых задач заключается в том , что студенты не являются

обученные грейдеры, которые приводят в пользу предоставления обратной связи про-

цесс как можно более простым. Учитывая широкие доказательства того, что для

многие задачи порядковым обратной связи проще обеспечить и более

надежнее , чем кардинальное обратной связи [5, 38, 29, 9], то там-

Фор желательно базовой экспертной классификации по порядковому обратной связи (например ,

"Проект X лучше , чем по проекту Y"). К сожалению, все экс-

Методы ствовавших для агрегируются отзывы сверстников в общую

оценка требуют , чтобы студенты обеспечить кардинальное обратную связь

(например , "Проект X является B-"). Кроме того, эффективность простой

методы для агрегирования обратной связи кардинальное, такие как средняя

старение, была поставлена под сомнение [7, 10, 30]. В то время как вероятностный

Методы машинного обучения в последнее время было предложено

решения этих задач [32], они по- прежнему сталкиваются с проблемой , что

студенты могут быть классификации на разных масштабах. Например,

студенты могут иметь предвзятое , что представляет собой B +

|  |
| --- |
| **Страница 2** |

основанный на университете они происходят. Эти весы могут

также может быть нелинейным как разница между А + и

А , не может быть такой же , как разница между C + и

С.

Для того, чтобы преодолеть проблемы кардинального обратной связи, мы представляем задачу порядкового экспертной классификации в данной статье. От

имеющие студенты дают порядковые оценки и не кардинальные

заявления как обратная связь, мы разгрузить проблему развиваю-

ИНГ шкалу от студента на алгоритм экспертной классификации.

Основные технические вклады этой бумаги лежат в вырожденном

разработка методов для порядкового экспертной классификации, где

Цель состоит в том, чтобы автоматически вывести общую оценку набора

заданий от порядкового обратной связи со стороны сверстников. Кроме того,

Вторая цель наших методов сделать вывод , насколько точно

каждый студент обеспечивает обратную связь, так что надежность классификации может

быть стимулы (например, в качестве компонента общей оценки).

Для этого, мы предлагаем несколько методов машинного обучения

для порядкового экспертной классификации, которые отличаются тем , как вероятность Отбрасывая

более рейтингах вклады моделируются. Для этих моделей мы

обеспечивают эффективные алгоритмы для оценки сортов назначения

и грейдер надежностей.

Изучить применимость наших методов в реальном мире

Настройки, мы собрали данные экспертной оценки как часть с университетским

Конечно , уровень. Используя эти данные, мы демонстрируют эффективность

предлагаемых порядковых методы обратной связи в сравнении

к существующим методам кардинальное обратной связи. Более того,

мы сравним наши порядковые методы равный с классификации тра-

методы оценки ные , которые были использованы в процессе в

параллельны друг другу. Используя эти данные в классе мы также исследовать другие

свойства этих методов, таких , как их надежность, данные

зависимость и самосогласование. Наконец, мы анализируем повторно

в ответах обследования , проведенного студентами в классе

эксперимент, показывая , что большинство студентов нашли сверстников

градация опыт (получение и предоставление обратной связи) help-

FUL и ценным.

**2. PEER СОРТИРУЯ ПРОБЛЕМА**

Мы начнем с формального определения задачи экспертной классификации,

поскольку он представляет себя с точки зрения машинного обучения.

Нам дан набор | D | назначения D = {d 1, ..., d | D | }

(например, эссе, отчеты) , которые должны быть этапным. градуировки

сделано с помощью набора | G | грейдеры G = {г 1, ..., г | G | } (Например, студент

равный грейдер, рецензенты), где каждый получает грейдер подмножество

D G ⊂ D для оценки. Выбор заданий для каждого грейдер

может быть равномерно случайным образом , или может следовать за детерминированный или

последовательный дизайн. В любом случае, количество назначений

что любая грейдер оценка | D г | намного меньше, чем общее

количество заданий | D | (например, | D г | ≈ 10).

Каждый грейдер обеспечивает обратную связь для его набора назначаемыми

менты D г. Порядковый и кардинальное сверстников классификации различаются в

тип обратной связи грейдер , как ожидается, даст:

Кардинал Peer Градация (CPG): В кардинального пэра постепен-

ING, каждый грейдер г обеспечивает кардинальное-значных обратной связи для

каждый элемент d ∈ D г. Как правило, это числовое или кошка-

egorical ответ , который мы обозначим через у (г)

d

(например, Ликерт

Шкала, письмо класс).

Порядковый Peer Градация (OPG): В порядкового экспертной классификации,

каждый грейдер г возвращает заказа а (г) (возможно , с

связи) его или ее назначения D г, что указывает на относительное , но

не абсолютное качество. В более общем смысле , порядковое обратной связи

может также состоять из нескольких попарных предпочтений, но

мы сосредотачиваемся на случае одного упорядочения в этой статье.

D г (⊂ D) Набор предметов рассортированы грейдер г

s d (∈)

Прогнозируемая оценка для элемента D ( чем больше , тем лучше)

η г (∈ +) Прогнозируемая надежность грейдер г

σ г

Оценка обратной связи (с возможными связями) от г

г (σ)

d

Ранг пункта й в рейтинге сг (ранг 1 лучше)

ρ г

Набор парного обратной связи предпочтительно от г

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

Таблица 1: Обзор Обозначения и ссылки.

Независимо от типа обратной связи , которые обеспечивают грейдеры,

цель в экспертной классификации имеет два аспекта.

Мы называем первую оценку цели класса, что задача

оценки истинного качества заданий в D из

обратная связь грейдер. Различают два типа

оценка сорта, которые отличаются тем , как они выражают присваивание

ния качества. По оценке порядкового класса, цель состоит в том, чтобы

вывод ранжирования сг всех заданий в D , что большинство точно-

дельно отражает некоторое истинное упорядочение (по качеству) а \*. В cardi-

Оценка NAL класс, цель состоит в том, чтобы вывести кардиналом Ранг S D

для каждого d ∈ D , что наиболее точно отражает каждый истинный класс

s \*

d. Обратите внимание , что тип обратной связи не обязательно DE-

Termine ли выход оценки сорта является порядковое или

Кардинал. В частности, мы увидим , что некоторые из наших методов

может вывести кардинальные оценки , даже если только дают порядковое обратную связь.

Вторая цель является оценка надежности грейдер, который

задача оценки , насколько точной обратной связи грейдера

является. Оценка надежности грейдер имеет важное значение , по крайней мере

две причины. Во- первых, выявление ненадежных оценки позволяет

чтобы понизить вес их обратной связи для оценки класса. Во- вторых,

и что еще более важно, это позволяет нам стимулировать хорошие и

тщательной классификации путем пэром классификации себя часть

общая оценка. В дальнейшем мы будем , как правило , представляют собой

надежность грейдера в качестве единственного числа п г ∈ +.

В следующих разделах мы выводим и методы оценки

для оценки класса и оценки надежности грейдера в

Порядковый установка Peer классификации.

**2.1 Связанные работы в ранге Aggregation**

Проблема оценки оценка в порядковом Peer Grading

можно рассматривать в качестве специфического типа проблемы агрегации ранга.

Ранг агрегации описывает класс проблем , связанных с

комбинируя информацию , содержащуюся в рейтинге с муль-

tiple источники. Многие популярные методы , используемые сегодня [15, 25, 11]

основываться на классических моделей и методов , таких как полугруппа

NAL работа Терстоун [39], мальвы [28], Брэдли и Терри

[8], Люс [27] и Плакетта [33]. Эти методы были

используется в различных областях, каждая из которых отпочковывались

их собственный набор методов.

Результат поиска Aggregation (также известный как ранг буду-

Sion или метапоисковой) имеет цель результата поиска присоединяемого

рейтинги из различных источников для получения одновыходовые

ранжирования. Такая агрегация широко используется для улучшения

за выполнением какого - либо одного Ranker в обоих под наблюдением

и неконтролируемые настройки [3, 34, 40, 31]. агрегация место

для поиска отличается от Порядковый Peer грейдов в нескольких AS-

сужаются. Во- первых, оценка надежности грейдер не цель

сам. Во- вторых, успех результата поиска агрегации Де-

в основном на сит правильно определить главные элементы, в то время как

оценка оценка направлена точно оценить полную rank-

ING. В- третьих, связи и разреженности данных не является проблемой в поиске

агрегирование результата, так как ( по крайней мере , в принципе) вход rank-

ь в общее количество заказов по всем результатам.

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

Алгоритм 1 Нормальный Кардинал-Score (NCS) Алгоритм

( так называемый PG 1 в [32]) используется в качестве базовой линии в наших экспериментах

s d ~ N (μ 0, 1

γ 0

)

⊳ Истинные результаты

η г ~ Gamma (α 0, β 0)

⊳ грейдер Надежность

б г ~ N (0, 1

γ 1

)

⊳ грейдер Bias (Только для NCS + G)

у (г)

d

~ N (s d + Ь г, 1

η г

)

⊳ Наблюдаемые Кардинал Peer Оценка

Расчетный показатель сек d, η г и б г

⊳ Использование ОМП

Социальный выбор и Системы голосования выполняют звание агрес-

купности предпочтений , что набор лиц заявили более

конкурирующие позиции / интересы / кандидатов. Цель состоит в том, чтобы иден-

тифы наиболее предпочтительные альтернативы , данные противоречивы цен предпочтительнее

почтений [2]. Обычно используемые методы агрегации являются

Борда рассчитывать и другие схемы голосования Кондорсе [3, 13, 26].

Эти методы плохо приспособлены для задачи OPG, так как они

не модель надежности избирателей, обычно предполагают ранжирование

все альтернативы (или , по крайней мере , оставить выбор альтернатив вверх

избирателю), и , как правило , сосредоточены на верхней части рейтинга.

Краудсорсинг, вероятно, наиболее тесно связанных с при-

домен пликация, где цель состоит в том, чтобы объединить обратную связь

из нескольких crowdworkers [19, 6]. В силу различий

Качество этих работников, моделирования надежности работника

существенным [35, 11]. Основное различие в наших условиях является то , что

количество элементов велико , и мы хотели бы, чтобы правильно

заказать все из них, а не только определить топ-мало.

Ранг-агрегация также используется для других параметров

такие как MultiLabel / мультиклассируют классификации (путем объединения

различные классификаторы) [23] или для обучения навыков игрока в азартным

ING окружающей среды [18]. Является ли невозможно обследовать обширную лите-

ратура по данной теме , и , таким образом , мы отсылаем заинтересованного читателя к

комплексное обследование по этой теме [24]. Эти методы

также были адаптированы к оценке образования [4], с помощью

графическая модель на основе подхода, для моделирования трудно-

вопросов , трудность и оценки правильных ответов

Crowdsourced установка. Однако эти методы не являются ни

применяется для настройки экспертной классификации они не могут справиться

неокончательный ответы (например , эссе).

**2.2 Связанные работы в Peer Grading**

С появлением онлайновых курсов, сверстников классификации было

все чаще используется для больших классов со смешанными результатами [7, 10,

30]. В то время как большинство предыдущих использование экспертной классификации имеют повторно

лгал на простых методов оценки , как в среднем количественной обратной связи, в последнее время вероятностный обучения алго-

rithms был предложен для оценки экспертной класса [32].

Тем не менее, этот метод требует , чтобы студенты обеспечивают автомобиль-

Динал баллы как классы. Второе ограничение метода

в работе [32] является то , что они надежность грейдер стимулирования, связывая

это собственный счет присвоения грейдера. Тем не менее, такое

установка неуместна , когда есть группы (такие , как наш

установка) или там , где внешние грейдеры / используются обозреватели (например,

конференции обзор). Кроме того, такой косвенный сти-

ный труднее общаться и оправдать по сравнению с

Оценки прямых грейдер надежности используется в нашем случае. наконец

их подход требует, чтобы каждый студент классов некоторые AS-

signments , которые ранее были классифицированы инструктором в

Для оценки надежности грейдер. Это кажется расточительным,

при условии , что студенты только в состоянии класса небольшое количество

Назначения в общей сложности. Мы опытным путем сравнить их кардинальных

равный метод классификации (алгоритм 1, с использованием ОМП вместо

Гиббс выборки) с методами классификации порядковое сверстников

Предложенная в настоящей работе.

В целом, учитывая ограниченное количество внимания , что

Проблема экспертной классификации получила в области машинного обучения

литература до сих пор, мы считаем , есть широкие возможности для

улучшить на внедренный и устранения недостатков

что в настоящее время существуют [36], которая усиливается одновременным

работа на тему другими [12, 37].

**3. порядковое МЕТОДЫ PEER СОРТИРОВКИ**

В этом разделе, мы разрабатываем методы профилирования порядковое одноранговые

для оценки класса , а затем распространить эти методы к

Проблема оценки надежности грейдер. Наши методы

публично доступны как программное обеспечение на www.peergrading.org,

где мы также предоставляем веб - сервис для экспертной класса осо-

timation. Эти методы требуют как данные н.о.р. образца

упорядочений

S = (σ (1 г), ..., σ (G | G |)),

(1)

где каждое упорядочение сортирует подмножество заданий в соответствии

на суд грейдер г I.

3.1 Оценка Оценка

Наши методы оценки класса основаны на моделях,

представляют распределения вероятностей над ранжировании. В частно-

улар, мы расширяем модель мальвы в (§ 3.1.1), то Bradley-

Терри модель (раздел 3.1.3), модель Терстоуна ( в п 3.1.4), и

модель Плакетта-Люс (раздел 3.1.5), которые необходимы для

порядковое проблема экспертной классификации.

3.1.1 Маллоус Модель (MAL и MALBC)

Модель мальвы в [28] описывает распределение по ранжировании

σ в терминах расстояния б (а, σ) от центральной Е рейтинга,

который в наших условиях является истинным ранжирование σ \* назначений

по качеству.

P (σ | a) =

е -δ (a, σ)

Σ

σ е -δ (a, σ)

(2)

Несмотря на то , учитывая соблюдение максимальной оценки вероятности а \*

рейтинга является NP-трудной для многих дистанционных функций [13, 34],

податливые приближения известны для особых случаев. В этом

работа, мы используем следующую сговорчивым расстояние Kendall-т

[20], которая предполагает , что оба рейтинга являются полными упорядочиваний

над всеми заданиями.

Определение 1. Определим Kendall-т расстояние б K Бе-

твин ранга сг 1 и 2 ранга , как сг

δ K (σ 1, σ 2) =

Σ

d 1 ≻ σ1 d 2

Я [[d 2 ≻ σ 2 d 1]]

(3)

Он измеряет количество неправильно упорядоченных пар между

два рейтинга. В нашем случае, рейтинги , что студенты

обеспечить может иметь связи. Мы рассматриваем эти связи как безразличия

(то есть, агностик либо ранжирования), что приводит к следующему

модель, где суммирование в числителе ведется по всем

всего упорядочиваний сг согласуется со слабой упорядочения а.

P (σ | a) =

Σ

σ ~σ е -δ (a, σ)

Σ

σ е -δ (a, σ)

(4)

Отметим также , что вход ранжирования сг может только своего рода подмножество

назначений. В таких случаях мы надлежащим образом Ограничим

Нормализация константа в (4). Для расстояния Kendall-т, этого

константа нормировки может быть вычислена эффективно, и это

зависит только от количества элементов в рейтинге.

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

|  |
| --- |
| **Страница 4** |

Алгоритм 2 Computing MLE ранжирования для мальвы модели

1: C ← D

⊳ C содержит неоцениваемый элементы

2: при г = 1 ... | D | делать

3:

для d ∈ C сделать

4:

х d ← Σ g∈G η г | d ∈ C: d ≻ σ г d | - | d ∈ C: d ≻ σ г d |

5:

d \* ← мин d∈C х d

⊳ Выберите самый высокий пункт подсчета очков

6:

г (σ)

d \*

← я

⊳ Позиция в следующий пункт

7:

C ← C / d \*

⊳ Удалить D \* из набора кандидатов

8: возвращение σ

Числитель может быть также вычислена эффективно. Заметка

что связи в грейдер рейтингах а (г) не влияют на обычный-

ризации постоянной при интерпретации равнодушия.

В соответствии с этой модифицированной модели мальвы, в максимальной как-

lihood оценкой сг центральной Ранжирование

Σ = σ Argmax

{

Π

g∈G

Σ

σ ~σ (г) е -δ K (σ, σ)

Z M (| D г |)

}

,

(5)

Вычисление оценки максимального правдоподобия сг как оце-

помощник истинного ранжирования по качеству а \* требует нахождения

Кемени-оптимальный агрегат, который , как известно, NP-трудной

[13]. Однако многочисленные приближения были изучены

в агрегации разряд литературы [13, 21, 1]. В этой работе

мы используем простой жадный алгоритм , как показано в алгоритме 2.

В качестве альтернативного алгоритма вычисления оцененный

рейтинга, мы используем Метод Борда-как приближение для

Модель Мальвы (которую мы обозначим , как MAL до н.э.), где линия 2

Алгоритма 2 заменяется

х d ← Σ

g∈G

г (σ (г))

d

,

3.1.2 Счет-Weighted Мальвы (MALS)

Модель мальвы представил выше , имеет два недостатка.

Во- первых, она не выводит содержательную кардинальное класс для

назначения, что делает его применимым только к порядковому класса

оценка. Во- вторых, расстояние б K не различает

между misordering заданий, которые похожи по качеству

от тех , которые имеют большую разницу качества.

Для решения этих двух недостатков, мы предлагаем продолжений

Sion , которая оценивает Кардинал сортах S д для всех заданий.

С этой целью мы вводим следующий счет-взвешенный

Ранжирование расстояние, которое масштабирует расстояние , индуцированное каждый

misranked пары по расчетной оценки разницы класса.

Определение 2. Счет-взвешенный Kendall-τ расстояние

δ SK над рейтинги сг 1, σ 2 , заданной Кардинал баллы S d является

δ SK (σ 1, σ 2 | s) = Σ

d 1 ≻ σ1 d 2

(с d 1 - d s 2) I [[d 2 ≻ σ 2 d 1]].

(6)

Лечение связей в грейдер рейтинге , как описано выше , повторно

таты в очковой-взвешенной версии модели мальвы (MALS).

Мы используем следующую максимальную апостериорной Оценщике осо-

леть счеты с.

s = s Argmax







Pr (s)

Π

g∈G

Σ

σ ~σ (г) ехр (-δ SK (σ, σ | s))

Σ

σ ∈π (D г)

ехр (-δ SK (σ, σ | s))







(7)

Заметим , что σ может быть получен путем сортировки предметов , как за единицу времени D.

Pr (s) = Π d∈D Pr (s d) является до на скрытые оценки позиций.

В наших экспериментах мы моделируем Pr (s d) ~n (0, 9), и использовать

то же самое перед во всех наших методов. В то время как в результате наблюдавшейся

проективное не обязательно выпуклый, мы используем стохастический Градиент

Спуск (SGD) для оценки класса и инициализации классов

используя уменьшенную решение мальвы.

3.1.3 Брэдли-Терри Модель (BT)

Приведенные выше модели определяют распределение над рейтинга в качестве

функция ранжирования расстояния, и они требуют приблизительная

методы решения проблемы максимального правдоподобия. В виде

альтернатива, мы можем использовать модели агрегации на основе ранга

на распределений по попарных предпочтений, поскольку ранжирования

N элементов можно также рассматривать как набор предпочтений за

(п

2) пары элемента. Модель Брэдли-Терри [8] одна модель

для парных предпочтений, и она получает распределение на основе

о различиях , лежащих в основе значения множества S через D

Логистическая функция связи.

P (d я ≻ р (г) d J | s) =

1

1 + е

- (S - ди - ди - джей -s)

(8)

Поскольку каждое решение предпочтение смоделированы в индивидуальном порядке ,

обратная связь от грейдера может быть (возможно , непоследовательно)

набор предпочтений , которые не обязательно должны образовывать

последовательное упорядочение. Ниже приводится ограничение максимальной дистальных

Априорные оценки используется в данной работе.

s = s Argmax







Pr (s)

Π

g∈G

Π

d я ≻

ρ (г)

d J

1

1 + е

- (S - ди - ди - джей -s)







(9)

Полученная задача (совместно) лог-выпуклым во всей очерков

ниваются сортов S d, причем градиенты принимает простую форму.

Следовательно синг может быть использовано для оценки глобального оптимального

сорта эффективно. Мы рассматриваем связи как отсутствие цен предпочтительнее

разностная. Можно также расширить эту модель , чтобы включить связи

более явно, но мы не будем обсуждать это для краткости.

3.1.4 Терстоуна Модель (ЧГ)

Заместитель логистическую функцию связи в Bradley-

Терри модель использовать нормальное распределение для спаривание

мудрые предпочтения. Как и модель Брэдли-Терри, Ре-

кающего Терстоуна модель [39] модель можно понимать как

случайная полезная модель , используя следующий процесс: Для

каждая пара элементов d я, d J образцы грейдер (латентное) Val-

ЕЭС х (г)

d я

~ N (s d я, 1

2) и х (г)

d J

~ N (s d J, 1

2), а затем заказы

пара на основе двух значений. Среднее значение нормали

распределение D I является качество сек d я. Ограничение максимальной posteri-

оценка ори партитур с требует максимизация

следующие функции:

s = s Argmax







Pr (s)

Π

g∈G

Π

d я ≻

ρ (г)

d J

F (s d я - s d J)







(10)

F является CDF стандартного нормального распределения. Эта

целевая функция логарифмически выпукло и мы используем для оптимизации SGD

Это.

3.1.5 Плакетта-Люс Модель (PL)

Недостатком попарных моделей предпочтение , что они

может быть менее выразительным , чем у моделей , построенных на распределениях

более рейтинга. Расширение модели Брэдли-Терри

(модель Плакетта-Люс [33]) позволяет использовать распределение

над ранжировании, сохраняя при этом выпуклость и простоту

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

Алгоритм 3 Переменный SGD на основе Минимизация

Требуют: N ≥ 0 (Количество итераций), правдоподобие L

1: Obj ← - журнал L

2: s ← SGD S (Obj, η = 1)

⊳ Est. оценки ж / о надежностей

3: для я = 1 ... N сделать

4:

η ← SGD G (Obj, s)

⊳ Расчетный показатель надежностей

5:

s ← SGD S (Obj, η)

⊳ Est. счеты с надежностей

6: возвращение s, η

градиентных вычислений. Эта модель может быть лучше всего можно понять

в качестве многоступенчатой опыт а где на каждом этапе, пункт d я

обращается (без замены) с вероятностью α е s - ди.

вероятность наблюдения ранжирования а (г) при этом процессе:

P (σ (г) | s) = Π

d я ∈D г

E S - ди /

(

E S - ди +

Σ

d я ≻

σ (г)

d J

е

s DJ

)

Полученная максимальная апостериорная оценка является

s = s Argmax







Pr (s)

Π

g∈G

Π

d я ∈D г

E S - ди

E S - ди + Σ

d я ≻

σ (г)

d J

е

s DJ







, (11)

**3.2 грейдера Оценка надежности**

В то время как методы , рассмотренные в разделе 3.1 , позволяют нам

оценки сортов назначения от порядкового обратной связи, они по- прежнему

не дают нам средства непосредственно оценить бульдозерным надежностей

Η г. Тем не менее, существует общий путь расширения всех методов

представленные выше , чтобы включить бульдозерным надежностей. С помощью

Модель Маллоус в качестве примера, мы можем ввести п г в виде

Параметр вариабельность следующим образом :

Pr (σ | a, η г) =

Σ

σ ~σ (г)

ехр (-η г δ K (a, σ))

Z M (η г, | D г |)

(12)

Полученная оценка обоих σ и п является

Σ, η = Argmax σ, η

{

Π

g∈G

Pr (η г)

Σ

σ ~σ (г) ехр (-η г δ K (σ, σ))

Z M (η г, | D г |)

}

, (13)

где Pr (η г) является до по надежности грейдера. В этом

Работа , которую мы используем Gamma перед п г ~ Gamma (10, 0,1).

Аналогичным образом , другие цели также могут быть расширены в

такой способ , как показано в таблице 2. В то время как многие из эксперимен-

имели тенденцию цели, такие , как один из приведенных выше в формуле. (13), являются

выпуклая в грейдер надежностей п г (для данного а), они ООН-

к счастью , не совместно выпуклостью в надежностей и

оцениваемые сорта. Таким образом , мы используем итеративный alternating-

способ минимизации, который чередует между минимиза-

ING в лог-цели для оценки классов и назначений

сведение к минимуму лог-цели для оценки грейдер надеж-

зательств. Этот итеративный подход с использованием переменного случайная по

крестики градиентного спуска используется для всех совместных задач оценки в

Эта бумага. Следует отметить , что методы которые оценивают в надеж-

ностями с использованием алгоритма 3 обозначаются A + G суффикс к

метод, в то время как те , которые просто оценивают назначение

сорта представлены только имя метода.

4. ЭКСПЕРИМЕНТЫ

Ниже мы приводим опыты, сравнивающие порядковые и кардинальное пэра методы классификации. Мы оцениваем их

способность предсказывать оценки инструктора, их изменчивость, их

Метод Оценка Cnvx оценщик

MAL + G

Нет

Нет Pr (η) Π g∈G

Σ

σ ~σ (г)

ехр (-η г δ K (σ, σ)) / Z M (η г, | D г |)

MALS + G Да

Нет Pr (иs, η) Π g∈G

Σ

σ ~σ (г)

ехр (-η г δ SK (σ (г), σ, F)) / Z (·)

BT + G

да

Да Pr (иs, η) Π g∈G

Π

di≻ρ (г) ди - джей

1 / (1 + е -ηg (ИПД-SDJ))

ЧГ + G Да

Да Pr (иs, η) Π g∈G

Π

di≻ρ (г) ди - джей

F (√η г (s - ди - ди - джей -s))

PL + G

да

Да Pr (иs, η) Π

g∈G

Π

di∈Dg

1 / (1+ Σ

di≻ρ (г) ди - джей

е -ηg (ИПД-SDJ))

Таблица 2: Резюме порядковыми методов изучены

какая модель надежностей грейдера, в том числе

Возможность вывода кардинальных баллов , и если в результате

Цель выпукло в этих баллов.

устойчивость к плохим сверстников классификации, а также их способность идентифицировать

плохие грейдеры. Мы также представляем результаты с качественной

опрос студентов , чтобы оценить , как студенты воспринимают пира

градация процесс.

**4.1 Сбор данных в классе эксперимента**

Мы используем реальный набор данных , состоящий из обратной связи со стороны сверстников, Т.А.

классов и классов инструктора для оценки сверстников градуировку

методы , предложенные в этой статье. Эти данные были собраны в качестве

часть старших студентов и мастеров уровня класса с

обучалось около 170 студентов. Класс был укомплектован

с 9 ассистентов (TA) , которые участвовали в оценивании, и один инструктор. Этот размер класса является привлекательным,

так как она была достаточно большой для сбора значительного количества

от сверстников классов, в то же время позволяя традиционным

инструктор и TA градация , чтобы служить в качестве основы. доступны

Способность инструктора классов делает наши данные отличаются от

другие оценки экспертной классификации , используемые в прошлом (например, [32]).

Мы рады предоставить данные для других исследователей суб-

ек к IRB утверждения.

Набор данных состоит из двух частей , которые были классифицированы без-

зависимым от, а именно презентация плаката и окончательный

доклад на длительный курс проекта 8 недель. Студенты работали

в группах по 3-4 студентов на протяжении всего проекта,

и там было в общей сложности 44 проектных групп. В то время как студент

работали в группах, равный градация была выполнена в индивидуальном порядке

с помощью набора инструментальных средств управления Microsoft Conference (СМТ)

система. Процесс сортировки равный проводили одно-

слеп для плакатов и двойных слепых для докладов и

рецензент задания были сделаны равномерно в случайном порядке.

Студенты были даны четкие указания и попросили , чтобы сосредоточиться на

такие аспекты, как новизна и ясности (среди прочих) в то время как De-

termining их класс. Они также просили , чтобы оправдать их

класс, предоставляя комментарии обратной связи. Студенты сказали

что часть их степени зависит от качества их

всмотреться обратной связи.

Все классификации было сделано на 10-балльной шкале (кардинал) Лайкерта,

где 10 был назван "идеальным", 8 "хорошо", 5 "пограничная", 3

"Дефицит" и 1 "неудовлетворительно". Это позволит нам COM-

стричь кардинал и порядковые методы профилирования сверстников, где OR-

Динал методы просто использовать порядок (возможно , со связями)

вытекает из кардинальных оценок. Обратите внимание , что в истинном приложениях

ние порядкового точности равный классификации может улучшить, так как

позволило бы упростить инструкции сортировки и сокращения

когнитивный над головой , если студенты не придется беспокоиться о том

точное значение конкретных кардинальных классов.

Ниже описываются процессы , используемые сортируя на каждом

этап, и таблице 3 приведены некоторые основные статистические данные.

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

Статистические данные

PO FR

Количество присвоений

42

44

Количество рецензентов 148 153

Всего экспертных обзоров

996 586

Всего TA Отзывы

78

88

Участие ТА

7

9

Per-Пункт Peer Оценка Devn. 1.16 1.03

Задавать

Кто?

Среднее Devn.

PO

Сверстники

8,16

1,31

TAs

7,46

1,41

Мета

7,55

1,53

FR

Сверстники

8,20

1,35

TAs

7,59

1,30

Инструктор 7,43

1.16

Таблица 3: Статистика для двух наборов данных (PO = Poster,

FR = Отчет) из классной комнаты эксперимента по

с персоналом (TAS / Meta / инструктор) и студента

распределения классов.

**4.1.1 Оценивание Процесс стендовых докладов**

Стендовых докладов проходило в два часа плаката

сессия. Две группы не представили свой плакат. Студенты

было предложено , чтобы повернуть их представления плакат. Эта

вероятно , увеличилась изменчивость сортов, так как разные Обзор -

ERS часто видели разные ведущие. Студенты и TAs приняли

ноты и вошли в свои отзывы с помощью СМТ впоследствии.

Сорта TA были независимыми, а это означает , что Тас

не видел экспертных обзоров перед входом их рассмотрения.

Были в среднем 1,85 обзорах ТП для каждого плаката.

В Peer Сорта составил в среднем 23,71 обзоров для каждого

постер, с каждым Рецензент обзором 6,73 плакаты на

в среднем.

Окончательный Мета Сорт для каждого плаката была определена как

следующим образом . Один из Тас , которые уже обеспечили неза-

вмятина обзор был выбран в качестве мета-обозревателем. Данный проект ТП был

предложено объединить все аргументы выдвинутые в

обзоры и сделать итоговую оценку по той же 10-балльной шкале.

Инструктор курировал этот процесс, но вмешался только

очень мало сортов.

**4.1.2 Процесс классификации для окончательного проектов**

В конце проекта, группы представили доклад

около 10 страниц. Процесс рецензирования был похож

к тому, что из стендовых докладов, но с одним важным

Разница - а именно, что все проектные отчеты были оценены ассистентами и инструктор без каких-либо знаний о

всмотреться отзывы, как описано ниже.

В среднем каждый отчет получил 13,32 оценок сокурсников как

Общий балл по каждому из обзоров (студенты были также

попросили компонентов оценки, как "ясности" и т.д.).

Каждый отчет также получил два TA Сорта, которые Тас

представленные без ведома экспертных обзоров.

И, наконец, каждый отчет получил инструктор Оценка, Следуя

традиционный процесс классификации проектов в этом классе.

Инструктор и руководитель ТП каждый сортовой половину проектов

и определили класс на основе их собственного чтения

бумага, принимая TA рассматривает в качестве входных данных. Эти сорта

были предоставлены без просмотра экспертных обзоров. Мы можем

Поэтому просматривать оценки инструктора как оценку, которая

совершенно не зависит от сортов одноранговых (в отличие от

Мета-классы для плакатов, которые имеют некоторую зависимость).

**4.2 Оценка Метрики**

Обычно используемая мера для отчетности работы студента

(среди многих стандартизированных тестов) является процентиль

Оценка по отношению ко всем студентам в классе. Следуя этому

практика, мы используем процентиля, как и сам класс (письмо

оценка может быть легко получены с помощью искривления), причисляя отчет

метрики как наши основные показатели деятельности. В частно-

лар, мы используем следующий вариант Кендалла-т, учитывающий

для связей.

**20,6**

**23,3**

**20,8**

**31,9**

**21,8**

**21,4**

**20,6**

**23,7**

**16,2**

**29,7**

**19,8**

**31,9**

**19,8**

**33,0**

**25.5**

**30,2**

**15**

**20**

**25**

**30**

**35**

**40**

**плакат**

**доклад**

**SCAVG**

**NCS**

**MAL**

**MALBC**

**MALS**

**BT**

**ЧГ**

**PL**

Рисунок 1: Сравнение методов экспертной классификации (з / о

грейдер оценка надежности) против Мета и Ин-

конструктору Оценки в терминах E K ( чем ниже , тем лучше).

τ KT (σ 1 , σ 2 ) =

Σ

d 1 ≻ о1 d 2

Я [[d 2 ≻ сг 2 d 1 ]] +

1

2

I [[d 1 ≈ сг 2 d 2 ]] (14)

Следует отметить, что эта мера не является симметричным, если предположить, что

Первый аргумент является целевой рейтинг и второй аргумент

является предсказанным ранжирование. Она относится к связи в целевом рейтинге

как безразличие. Связи в предсказанной рейтинге лечат

как отсутствие информации, подвергаясь 1

2 ошибки (то есть, эквивалентная

к разрыву связей в случайном порядке). Такая коррекция необходима

для целей оценки, так как в противном случае предсказывали рейтинги

не со всеми связей (которые не передают никакой информации) повлечет за собой никаких

ошибка. Нормализация τ KT (σ 1 , σ 2 ) и учитывая то ,

что мы можем иметь более одной цели ранжирования приводит к

следующая мера ошибок.

Определение 3. Принимая во внимание набор целевых ранжирование S г , мы раз-

оштрафовать ошибки Kendall-τ E K предсказанного рейтинга сг Я , как:

E K (σ I ) =

100

| S г |

Σ

σ т ∈S г

τ KT (σ т , σ I )

Max σ∈π (D) τ KT (σ т , σ)

(15)

Эта ошибка макро усредняет (нормализованная) Т KT ошибки для

каждая цель ранжирования. Благодаря нормализации, они лежат Бе-

твин 0 (что указывает на совершенное соглашение) и 100% (с указанием

обращение с целевыми ранжировании). Случайный рейтинг имеет экс

ожидаемому E K погрешность 50%.

**4.3 Насколько хорошо порядковое против кардинала Peer**

**СОРТИРУЯ Предсказать Итоговая оценка?**

Первый вопрос, который мы рассматриваем в том, как далеко сверстников классификации

напоминает баллы, инструктора. В частности, мы

исследовать ли методы достижения порядковое экспертной классификации

аналогичные показатели в качестве методов кардинальное сверстников классификации, даже

хотя порядковые методы получают строго меньше информации.

Для всех методов, рассмотренных в данной статье, Рисунок 1 показывает

ошибка Kendall-τ E K по сравнению с Сортах мета для

Плакаты, и по сравнению с Сортах инструктором

Отчеты. В errorbars показывают стандартное отклонение на

используя бутстраповский типа передискретизации.

На плакатах, ни один из методов существенно не показывают

хуже производительность, чем другой метод. В частности,

нет никаких доказательств того, что методы кардинальных работоспособнее лучше, чем порядковые методов. Аналогичный вывод

также справедливо и для отчетов. Тем не менее, здесь порядковое ме-

ODS на основе модели мальвы Выполним лучше, чем тонер

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

**23,6**

**30,0**

**23.5**

**31,3**

**27.5**

**29,3**

**25,2**

**30,7**

**22,2**

**34,3**

**23,9**

**31,4**

**24,2**

**31.5**

**27,3**

**31,9**

**15**

**20**

**25**

**30**

**35**

**40**

**плакат**

**доклад**

**SCAVG**

**NCS**

**MAL**

**MALBC**

**MALS**

**BT**

**ЧГ**

**PL**

Рисунок 2: Сравнение методов экспертной классификации (з / о

Надежность грейдер оценка) в отношении классов ТП в

условия E K , используя оценки ТП в качестве целевого рейтинга.

Динал NCS 1 метод [32] (см алгоритм 1), а также некоторые

из других порядковых методов. Просто в среднем кардинал

десятки одноранговых классов, которые мы называем Score Усреднение

(SCAVG), выполняет на удивление хорошо.

Таким образом, большинство методов достижения E K от 20%

и 30% от обеих проблем, но все они имеют большой стандартным,

viations. E K появляется ниже для плакатов , чем для

проекты, которые можно объяснить тем фактом, что Meta

Оценка была под влиянием оценок сверстников. Но насколько хорошо

Е К от 20% до 30%?

**4.4 Как Peer Градация По сравнению с TA**

**Градация?**

Рассмотрим теперь, как Peer Градация сравнивает к наличию

каждое задание рассортированы ТА. Для классов средних,

TA градация по-прежнему может быть осуществимо. Поэтому интересно

знать, если TA градация явно предпочтительнее Peer СОРТИРУЯ

когда это возможно. Но еще важнее то, что между судьей

Соглашение между несколькими ТП может дать нам ориентиры

за точность Peer грейдов.

В качестве первой опорной точки, мы оцениваем, насколько хорошо Т.А.

Оценки отражают Сорта Мета для плакатов и Ин-

конструктору Оценки для отчетов. В частности, мы рассмотрим

градация процесс, где каждое задание оценивается по одной

TA, которая назначает кардинальную класс. Каждая сорта TA Фракцию

из заданий, и окончательное ранжирование заданий является

Затем вычисляется путем перебора всех кардинальных классов. Мы называем это

градация процесс TA классификации.

Мы можем оценить Е К ТА градуировку с Meta

Оценки и классы инструкторов, так как у нас есть несколько

классы ТП для большинства заданий. Мы случайным образом ресэмплировать

сорт TA из доступных классов для каждого задания,

вычислить рейтинг, а затем оценить среднее значение и стандартное

отклонение Е К более 5000 образцов. Это приводит к среднему

E K 22,0 ± 16,0 для плакатов и 22,2 ± 6,8 для

отчеты. Сравнивая их с Е К одноранговых классификации

методы на рисунке 1, мы видим, что они сопоставимы с

выполнение многих методов экспертной классификации - несмотря на то,

Е К ТА градуировки положительно предвзятым. Обратите внимание , что Meta

Оценки и классы инструкторов были назначены на основе

Те же сорта TA мы оцениваем против.

1 Мы настроили гиперпараметры модели NCS к max-

imize производительность. Мы также использовали фиксированную надежность грейдер

параметр в модели NCS, так как он обеспечивает более Пер-

чем при производительности таких оценки надежности (NCS + G).

**24,4**

**25,7**

**25,8**

**31,4**

**25,8**

**27,7**

**19,7**

**23,9**

**31,2**

**26,1**

**30,7**

**25,2**

**31,6**

**30,2**

**31,8**

**24,9**

**15**

**20**

**25**

**30**

**35**

**40**

**плакат**

**доклад**

**SCAVG**

**NCS**

**MAL**

**MALBC**

**MALS**

**BT**

**ЧГ**

**PL**

Рисунок 3: Самосогласованность методов экспертной классификации

(без оценки надежности грейдер) в терминах E K .

Чтобы избежать этого смещения и обеспечить более справедливое сравнение с Т.А.

классификации, мы также исследовали, насколько последовательны сверстников сорта

с оценками ТП, а также, насколько последовательны сортов ТП на ходятся

промежуточные кадры различных ТСУ. На рисунке 2 показана E K равноправного узла постепен-

методы ING при использовании ТП Сорта качестве целевого ранжирования для

как плакаты и отчеты. Расхождения были снова оце-

повязана с помощью начальной загрузки ресэмпинга. Обратите внимание, что TA Оценки были

представленные без знания одноранговых классов. В общем и целом,

сорта сверстников имеют E K с оценками TA , что является SIM-

ИЛАР к E K с соответствующими Итоговые оценки рассматриваемых

в предыдущем пункте. Опять же, нет никаких доказательств того, что

порядковые методы экспертной классификации являются менее предсказанием из

TA классов, чем методы классификации кардинальное сверстников.

Для оценки E K между различными TAS, мы используем следу-

мычание процедура передискретизации. В отпуск-один из моды,

мы рассматриваем сорта случайно выбранного ТП в качестве мишени

получить рейтинг и вычислить предсказанную ранжирование по выборке

от других сортов ТП, как это описано выше. осреднение

более 5000 повторений показывает , что Е K между Тас

47,5 ± 21,0 для плакатов и 34,0 ± 13,8 для отчетов.

Эти цифры можно сравнить с Е К пэра постепен-

методов на рисунке 2. Для получения отчетов о Инг, сверстников сорта

примерно как в соответствии с оценками ТП как другие марки TA

находятся. Для плакатов методы экспертной классификации являются существен-

более предсказанием существу сортов TA, чем другие марки ТП.

Причиной этого является то, по крайней мере, в два раза. Во-первых, равный постепен-

методы ИНГ имеют доступ к гораздо большему количеству данных, что снижает

Изменчивость (тем более, что презентации не всегда

дается той же студента). Во-вторых, равноправный классификации мето-

ODS достаточно данных для коррекции различных оценочных шкал,

в то время как смещения в классификации масштабах может иметь катастрофические последст

в TA ствия классификации.

Наконец, мы также рассмотрим самосогласованность партнёру

классификации методов. По аналогии с самосогласованности Т.А.

градация, мы спрашиваем, насколько похожи являются сорта мы получим,

повторить процедуру классификации с другим образцом ассо-

sessments. Мы случайным образом разделить рецензентов на две части

одинакового размера наборов данных. Для каждого метода экспертной классификации, мы

выполнять оценку класса на обоих наборах данных, который генерирует

два ранжирование заданий. Связи в этих рейтингах являются

сломаны случайным образом, чтобы получить общее количество упорядоченности. Рисунок 3 показывает

E K между двумя ранжировании (более 20 разделов выборочные).

Для плакатов, равный градация значительно больше уверенности в со-

последовательны, чем ТП классификации, а также для отчетов всех сверстников классификации

методы имеют более низкие E K оценки , чем TA классификации , а также.

|  |
| --- |
| **Страница 8** |

15

20

25

30

35

40

25%

50%

75%

100%

SCAVG

NCS

MAL

MALBC

MALS

BT

ЧГ

PL

15

20

25

30

35

40

25%

50%

75%

100%

15

20

25

30

35

40

2

3

4

15

20

25

30

35

40

2

3

4

5

6

7

Рисунок 4: Изменение E К производительности методов экспертной классификации ( с использованием мета и инструктор Сорта цели

ранжирование) при изменении количества заданий, назначенных каждому рецензенту для постеров (первый слева) и

Отчеты (второй), а также при изменении числа рецензентов для постеров (третьи) Отчеты (последний).

В целом, мы приходим к выводу, что нет никаких доказательств того, что TA

градация привело бы к более точным результатам классификации

чем сверстников классификации.

**4.5 Как Градация точность масштаба с**

**числа экспертных оценок?**

Сколько рецензенты необходимы для точного взаимного постепен-

ING, и сколько мнений делает каждый узел грейдер нужно делать?

Для того, чтобы оценить, как изменения производительности с числом сверстников

обзоры, мы провели две серии экспериментов. Сначала мы

создано 20 более мелких наборов данных посредством субдискретизации количество

рецензенты. Результаты показаны в двумя крайними правыми

графики Рисунок 4. В целом методы ухудшать грациозно

когда число рецензентов снижается. В целом, мы находим

что большинство порядковые методы масштабирования, а также методы кардинальных

для обоих наборов данных.

Второй способ увеличения или уменьшения количества

Имеющиеся данные заключается в количестве заданий, что каждый

оценки студентов. Таким образом, мы повторили эксперимент, но не-

вместо него субдискретизации количество заданий на рецензента

(Что соответствует более низкой рабочей нагрузки для каждого грейдер).

Левые два участка рисунке 4 показаны результаты. Опять же, мы

считают, что производительность падает изящно.

**4.6 Могут ли методы Peer Оценивание Определение**

**Ненадежные Грейдеры?**

Peer градация может работать только на практике, если грейдеры достаточно стимулы, чтобы сообщить точную оценку. Эта

может быть достигнуто путем предоставления сорта также для качества из

градация. В дальнейшем, мы исследуем ли грейдера

надежности оценок, предложенных в разделе 3.2 можно определить

классников, которые не являются прилежными.

Для обоих плакатов и проектов, мы добавим 10 "ленивых"

сверстников грейдеры, которые сообщают случайные оценки, сделанные из нор

MAL распределение которого среднее значение и дисперсия совпадает с

остальные грейдеров 2 . Для порядковых методов, это приводит к

в случайном оценке. Затем мы применяем пэра градуировку методы, оценивая соответствующие параметры надежности п г для

каждый грейдер с помощью 10 итераций переменного оптимизацию

Алгоритм. Мы тогда ранг грейдеров по их оценкам,

η г .

На рисунке 5 (вверху) показывает процент ленивых грейдеров, что

Оценка среди 20 классов с наименьшим п г . Ошибка

Столбики показывают стандартную ошибку более 50 повторных опытов с различными

ленивые грейдеры пробы. Большинство порядковые методы значительно

2 В противном случае было бы легко идентифицировать эти грейдеры.

**36.6**

**40,7**

**100**

**72,4**

**100**

**80,0**

**54,6**

**55,5**

**51,7**

**46,9**

**49,4**

**48,8**

**37,6**

**49,5**

**15**

**25**

**35**

**45**

**55**

**65**

**75**

**85**

**95**

**плакат**

**доклад**

**NCS + G**

**MAL + G**

**MALBC + G**

**MALS + G**

**BT + G**

**ЧГ + G**

**PL + G**

**С грейдер НАДЕЖНОСТИ**

**ОЦЕНКА**

**42,4**

**29,6**

**42,2**

**21,1**

**37,4**

**24,3**

**44,2**

**30,0**

**39,6**

**18,8**

**42,9**

**20,0**

**43,0**

**18,8**

**42,6**

**22,8**

**15**

**25**

**35**

**45**

**55**

**65**

**75**

**85**

**95**

**плакат**

**доклад**

**SCAVG**

**NCS**

**MAL**

**MALBC**

**MALS**

**BT**

**ЧГ**

**PL**

**BASELINE ЭВРИСТИЧЕСКИЙ**

Рисунок 5: Процент времени грейдер, который хаотичности

шиеся оценки и заказы заданий является одним из

20 не менее надежные классники (то есть, нижняя 12,5%).

опережают кардинальную метод NCS для обоих плакатов

и доклады. Варианты модели мальвы Выполним

очень хорошо, определяя около 70-80% от ленивых грейдеров

отчеты и все 10 ленивые грейдеры для плакатов.

более высокую производительность для плакатов, чем для докладов было

следовало ожидать, так как студенты обеспечивают 7 вместо 4-х классов.

На рисунке 5 (снизу) показаны результаты эвристического base-

линия. Здесь оценка оценка без оценки надежности

выполняется, а затем грейдеры ранжируются по их E K с

по оценкам рейтинга σ. Почти для всех методов, это производительность

хуже, ясно показывая, что оценка надежности является

выше в определении ленивые грейдеры. Мы находим аналогичные результаты

даже тогда, когда есть 100 + ленивые грейдеры, и мы исследуем

Надежность в следующем разделе.

|  |
| --- |
| **Page 9** |

-6

-4

-2

0

2

4

6

8

10

50

100

SCAVG

NCS + G

MAL + G

MALBC + G

MALS + G

BT + G

ЧГ + G

PL + G

-6

-4

-2

0

2

4

6

8

10

50

100

Рисунок 6: Изменение E K ( с помощью инструктора и Meta

Оценки в качестве целевого рейтинга) для (слева) Плакаты и

(Справа) Заключительный отчеты с добавлением входов

биговки количество ленивых грейдеров т.е. E K (с ленивой) -

E K (без ленивых). Отрицательное значение указывает на то, что

повышает производительность при добавлении этого шума.

метод

Плакаты

Отчеты

время выполнения

Runtime (+ G) Продолжительность времени выполнения (+ G)

NCS

0,32 ± 0,03

7,0 ± 0,55

0,20 ± 0,03

4,6 ± 0,25

MAL

0,01 ± 0,00

6,1 ± 0,11

0,01 ± 0,00

2,5 ± 0,03

MAL до н.э.

0,01 ± 0,00

5,1 ± 0,08

0,01 ± 0,00

2,5 ± 0,03

MALS 151,4 ± 12,39

418,7 ± 9,10

2,0 ± 0,13

4,2 ± 0,16

BT

0,46 ± 0,06

5,6 ± 0,38

0,21 ± 0,02

2,2 ± 0,10

ЧГ

57,9 ± 0,76

490,1 ± 7,45

12,2 ± 0,86

120,8 ± 1,03

PL

0,36 ± 0,03

4,2 ± 0,08

0,18 ± 0,01

2,0 ± 0,10

Таблица 4: Среднее время автономной работы (с и без грейдера

оценка надежности) и их стандартное отклонение

различных методов в CPU секунд.

**4.7 Как Robust являются Peer Градация методы ленивым Автогрейдеры?**

В то время как раздел 4.6 показал, что оценка надежности в порядковом сверстников классификации хорошо подходит для идентификации ленивые грейдеры,

мы также хотели бы знать, какой эффект эти ленивые восьмиклассника

на производительность оценки класса. Мы изучаем RO-

bustness методов экспертной классификации путем добавления возрастающие

ING количество ленивых грейдеров. На рисунке 6 показано изменение

E K (WRT Инструктор / Meta классы) после добавления 10/50/100

ленивые грейдеры ( по сравнению с E K с не ленивых грейдеров).

Мы считаем, что в большинстве случаев производительность не меняется

много по отношению к изменчивости методов. Интерес-

ingly, в некоторых случаях также улучшает производительность при добавлении

этот шум. Более глубокий осмотр показывает, что шум является наиболее

полезно для методов , чьи оригинальные E K выступление было

слабее, чем у других методов. Так , например,

Терстоуна модель показала самую слабую производительность по маршруту

Отчеты и улучшает самое.

**4.8 Как вычислительно эффективных являются Peer**

**Градация методы?**

В то время как точность предсказания является главной заботой класса

умозаключение, вычислительная эффективность должна быть достаточной, как

Что ж. В таблице 4 показаны средние среды выполнения и их уровень

отклонения для плакатов и докладов. Все методы

являются послушным и наиболее закончить в течение нескольких секунд. Счет-

Взвешенная модель Маллоус менее эффективен для задач, где

каждый грейдер оценивает множество заданий, так как градиенты

лор вычисления включает в себя вычисление нормализации со-

постоянен (который включает в себя суммирование по всем рейтингам). Однако,

обучение масштабируется линейно с числом классов. Другая

метод, который требует больше времени является модель Терстоуна.

Вопрос А) получал пэром

обратная связь полезна?

Вопрос Б) обеспечение

обратная связь сверстников ценным?

1 Да, это было полезно.

B 1 Да это был ценный опыт

2 Полезный, но не столько , сколько

инструктор обратной связи.

B 2 Да, это было ценным, но с

предостережений (например, взял много времени).

3 Немного полезно (например , только

Несколько комментариев были полезны).

B 3 Только мало значения (например , слишком Трудноудаляемые

ficult / не хватало навыков сортируя)

4 Нет / Не совсем / не имели

очень помогает.

B 4 Не ценный / Не очень ценной

состоянии.

5 Другое / Missing

B 5 Прочие / Missing

Таблица 5: варианты ответов на вопросы анкеты.

Основным узким местом здесь является вычисление градиен-

лор, как она включает в себя поиск по значению CDF из нормального

таблица распределения.

**4.9 Do Студенты Значение Peer классификации?**

Последний момент, который мы хотели бы изучить, что сверстников

градация не только о оценке сорта, но и о

генерирования полезной обратной связи. В частности, кардинал или

порядковые оценки были лишь небольшая часть сверстников запитке

назад. Peer грейдеры пришлось написать обоснование для их AS-

sessment и оставлять комментарии о работе в более общем плане.

Для оценки этого аспекта экспертной классификации, опрос был со-

Канальные в конце класса в рамках обратной связи курса про-

цесс. Данное исследование было включено два вопроса по поводу учащегося

всмотреться классификации опыт работы в классе; Говоря более конкретно, о

насколько полезной обратной связи они получили было, и как ценным

в состоянии опыт обеспечения обратной связи был для них. И то и другое

вопросы были даны ответы в свободной форме текста. Из 161

студенты, которые участвовали в проекте, 120 студентов повторно

откликался, по крайней мере, на один из вопросов, с 119 ответа

вопрос о получении обратной связи (средняя длина ответа

в символах: 62,93; STDEV: 77,22) и 118 вопрос

об обеспечении обратной связи (среднее значение: 100,36; STDEV: 105,74).

Следуя стандартной практике из анализа обследований, мы твор-

ованные пять категорий для кодирования этих открытые ответы

как показано в таблице 5. В то время как первые четыре категории (примерно)

следовать убывающую шкалу утверждения, последний служит в качестве

поймать-все (в том числе отсутствующих ответов).

Все свободно текстовые ответы были вручную назначены на эти

Категории четырьмя внешними аннотаторами (которые не были In-

, затронутый с классом и имел в своем распоряжении не видел комментарии).

Для всех 237 студентов комментариев (т.е., ответы), ангар-

notators было предложено выбрать категорию, которая больше всего

уместно / лучше всего описывает свой комментарий. Для проверки меж-

аннотатор соглашение, которое мы использовали меру Фляйсс Каппа. κ

значения 0.8389 и 0.6493 для двух вопросов указывают

высокая степень согласия аннотатор. Окончательное назначение ответа

к категории было сделано большинством голосов среди четырех Ан-

notators (оценка 0.5 каждый, если связаны между категориями).

В таблице 6 приведены результаты обследования после того, как Режим кодирования

ING. В целом около 68% сочли, по крайней мере до некоторой степени полезным

получать обратную связь со стороны сверстников, и около 74% нашли существен-

Значение циал в обеспечении обратной связи со стороны сверстников. Интересно, что из

26% студентов, которые выразили, что получение пэром

обратная связь не было (на самом деле) полезно для них, 17% до сих пор найдено

его ценным для обеспечения обратной связи со стороны сверстников. В целом, мы приходим к выводу

что подавляющее большинство студентов нашли какую-то ценность в

всмотреться классификации процесса.

**5. ВЫВОДЫ**

В данной работе мы исследуем проблему оценки студента на

Шкала с помощью экспертной классификации с использованием порядкового обратной связи. Мы бросили это как

проблема агрегации ранга и рассматривать различные вероятностные

|  |
| --- |
| **Page 10** |

%

1

2

3

4

5

Всего

B 1

34.58 2.08

5,83

10.00 1.67

54.17

B 2

5,42

0.00

5,83

7,08

1,67

20.00

B 3

0,42

2,92

2,08

2.50

0,42

8,33

B 4

2,92

0,83

5.00

5,42

0.00

14.17

B 5

0.00

0.00

0,42

1,67

1,25

3,33

Итого 5,83 19,17 43,33 26,67 5,00

Таблица 6: Результаты опроса учащихся, закодированы Ac-

ствии с категориями в таблице 5.

модели для получения оценки студентов, а также оценки

надежность одноранговых классов. Используя данные, собранные

реальный курс, мы видим, что производительность порядкового сверстников

классификации методов, по крайней мере конкурировать с кардинальным мето-

бам для оценки класса, несмотря на то, что они требуют строго

меньше информации от грейдеров. Для грейдера оценка надежности

, модель Маллоус в превосходит все другие методы,

и он показывает стабильно хорошие и надежную производительность

оценка класса, а также. В общем, мы находим, что порядковое

равный градация является надежной и масштабируемой, предлагая градацию точно-

колоритный, что сопоставимо с TA классификации в нашем курсе.