Использование согласованности разметок для статистической оценки качества рубрикатора

Федоров Артем Максимович

October 2024

Аннотация

Данная статья предлагает новый подход к анализу качества рубрикаторов, независящий от домена рассматриваемых объектов. Выдвигается предположение о мультимодальной природе распределения объектов рубрик в латентных пространствах современных моделей-векторизаторов на основе механизма внимания и предлагается пайплайн решения таких задач через восстановление смесей распределений в структурированном пространстве. Выводятся оценки разделимости распределений смесей von-Mises-Fisher семейства и проводятся статистические выкладки их устойчивости.

1 Вступление

В различных задачах рубриктор может носить различные имена: тематизатор для задач подразделения текстовых массивов по критерию их содержания [12, 18], ценностный или культурный код для описания систем ценностей в языке, классификатор для задач дискриминатного анализа в машинном обучении. В общем случае задача создания рубрикатора подразумевает под собой построение системы классов для некторой категории объектов, что могла бы приблизить ее природную структуру, при помощи которого объект мог бы быть описан совокупностью рубрик, к которым он принадлежит.

Тем не менее, основной проблемой создания рубрикатора является его потимальность: из прикладных целей ожидается, что рубрики будут между собой непересекающимися, легко различимыми и полными по своей наполненности. Первые работы на тему создания оптимального рубрикатора приводились в начале 2000 годов при изучении сложных объектов, требующих соответствуенно сложную стректуру классов для их описания, таких как язык и категоризация текстовых массивов [13, 28, 2]. Такие работы основывались на мнении экспертов соответсвующей области [13, 40], либо на применении простейших алгоритмов машинного обучения (LDA [34, 35] или же SVD [28]).

Тем не менее с развитием машинного обучения и все большим его применением в новых областях, создание рубрикатора стало применяться в качестве составной части конечного задания или решения, где полагаться на конструктивный подход организации рубрикатора стало невозможно либо из семантики задачи (автоматическое создание рубрикатора), либо по причине потребности в валидации его качества в бизнес целях. В данном контексте оценка качества должна быть нацелена на тесную работу с экспертами области для доведения руббрикатора до оптимальной в контексте задачи формы.

Несмотря на то, что данная область машинного обучения появилась на заре образования nlp, большинство методов получения мер качесвте рубрикаторов являются слишком невыразительными: если они и способны показать конечную метрику, то под возможность дать взглянуть на саму проблему при ее наличии и проанализировать ее – они не заточены. Однако возможность получить качесвтенный анализ получаемого рубрикатора является важной составляющей создания заданий для крауд сорс платформ (как составить интуитивно понятное задания для разметчиков) или в задачах автоматического построения рубрикаторов, для анализа уместности используемых подходов.

Данная работа является результатом решения потребности лаборатории Семантического анализа текстов Московского Государственного Университета им. М. В. Ломоносова в новой универсальной метрике качества рубрикатора, применяемом в проектах по разметке корпусов текстов, требующих сложные и большие по количеству системы рубрик (для разметки ценностей, эмоций, поляризации мнений в текстах и т. д.), позволяющей анализировать набор рубрик, найти в нем повторяющие друг друга элементы, выявить элементы, которых недостает.

2 Связанная работа

Задача оценки качества рубрикатора возникает в первую очередь в областях машинного обучения, где создание структуры классов является составной частью решения. К таким задачам в первую очередь относятся пайплайны крайдсорсинга и применения активного обучения [38, 25], где она выступает в качестве одной из ступеней валидации, в выявлении скрытой структуры данных через тематическое моделирование, [41, 43], где она сводится к оценке результата unsupervised или semi-supervised методов.

Методы оценки качества рубрикатора конструктивно строятся на векторизации данных. Классическим способом например получения аналитического

представления текстовых данных считается TF-IDF, широко применяемый на заре рубрикации данных [43, 42], применение методов PCA [21] и ICA [20]. Появляение же в последнее время сильных моделей векторизации данных различных доменов (текст, фото, звук, видео) ввиду развития нейросетевого подхода позволило приблизить решение задач компьютером к уровню разметчиков людей, позволяя рассматривать задачу оценки качества рубрикатора с изначальных пространств объектов (текст, видео, прочая медиа) на изучение многообразия объектов в латентном пространстве модели (будь то BERT [19] или clip [29] поднобный энкодер).

2.1 Особенности областей применения

В автоматической рубрикации данных [43, 39] оценка качества результата носит валидационный характер. В таких задачах постановка либо включает в себя предположительный набор рубрик с малым числом представителей, либо подразумевает выявление структуры классов с нуля с последующей передачей результатов разметчикам.

В задачах активного обучения и краудсорсинга основополагающим требованием является эффективность работы разметчиков на платформе [38], что диктуется прикладной областью (стоимость разметки для бизнеса и исследователей). Создание хорошо понимаемого и эффективно решаемого разметчиком задания является первой ступенью для удовлетворения данного требования. В таком случае, оценка рубрикатора классов, используемого разметчиками при обработке примеров, отражает согласованность разметчиков, насколько каждый из них пособен понимать устройство рубрикатора и определять соответсвие объектов предложенными рубриками. Задачи связанные с активной разметкой все чаще начали допускать разметчиков к самостаятельному определению множества рубрик, что наблюдается в исследовательских задачах при анализе работы LLM [17, 14, 9, 6], где оценка качества применяется на размеченных данных, или же ценностной наполненности языков.

2.2 Обзор известных методов

Такие задачи классически решаются методами восстановления класетрной структуры объектов, таких как k-Means [31, 22, 26], иерархическая кластеризация (чаще всего с расстоянием Уорда, максимальным или минимальным между подмножествами объектов) [23, 36], либо же алгоритмы восстановления плотностей в латентном пространстве, такие как MeanShift [8, 7] и DBSCAN [11]. Для первого подхода постановки задачи определения рубрикатора с semi-supervised данными, где для подмножества рубрик известны характерные представители, качество рубрикатора оценивается простейшими внешними метриками классификации PrecisionBCubed, RecallBCubed и так далее. В то же время для unsupervised подхода используются внутренние меры качества: кофенетический коэффициент корреляции (CPCC), индексы следа, Калинского или Гарабача [3, 24]

2.3 Противопоставление нашего подхода

Каждый из затронутых ранее подходов рассматривает векторизованные объекты как сгенерированную выборку из некоторой смеси распределений, каждая компонента которой соответсвует одной рубрике. Такой подход позволяет упростить модель данных в латентном пространстве и открывает путь к применению классических метрик качества для класетризации. Тем не менее утакого подхода есть ряд недостатков:

1. Предположение, что рубрики соответсвуют одномодальным распределениям, сильно ограничивает выразительность используемой модели.
2. Информация о том, что объекты были соотнесены разным рубрикам в процессе разметки используется при оценке качества рубрикатора не напрямую, но лишь является одним из параметров метода кластеризации, что приводит к невыразительности суждения о качетсве рубрикатора в целом.

Изучая устройство размеченных выборок для задач определения культурных ценностей, эмоций и полярностей мнений, мы имеем основания полагать, что распределения рубрик являются мультимодальными. И проведение кластеризации с последующим применением метрик качества не способно в действительности описать качество рубрикатора.

Наш подход предлагает новый взгляд. Мы считаем, что качественно обученный векторизатор способен хорошо описывать структуру объектов, позволяя оценивать уверенность разделимости рубрик на объекте мерой разделимости двух соответсвующих распределений в пространстве эмбеддингов. Тогда по размеченной выборке объектов мы способны восстановить кластерную структуру каждого класса при помощи EM алоритма [4], получить оценки максимального правдоподобия на параметры распределений и далее высчитать оценки разделимостей смесей, что можно интерпретировать в статистическом анализе.

Данный подход является логическим продолжением ROC анализа [27] в классической литературе по SignalDetection [16], и предполагает вывод нового индекса разделимости, ведущего себя как d' [30, 37], но учитывающего моменты до второго порядка распределений.

3 Методика алгоритма

Мы ограничиваем наш спектр задач рассмотрением ситуации, где эксперты уже смогли составить рубрикатор, по которому разметчики смогли составить размеченную выборку.

Как было предложено ранее, мы будем проводить анализ рубрикатора в три этапа:

1. Получение векторных представлений для объектов из обучающей выборки.

2. Восстановление аналитического распределения данных объектов в пространстве.

3. Расчет попарных оценок "различимости"(глава 4) классов с последующим их анализом.

3.1 Векторизация объектов обучающей выборки

Считаем, что каждый объект в общем случае помечается подмножеством меток из рубрикатора (задача multiclass classification вложима в задачу multilabel-classification) $x \sim \{0,1\}^{|T|}$ – кодировка. Положим независимость в совокупности всех классов $\Rightarrow$ для каждого класса $t \in T$ соберем коллецию из всех векторизованных представлений объектов, что были помечены данной меткой. Тем самым переходим к анализу $|T|$ коллекций объектов

3.2 Восстановление распределений векторов

Следующей глобальной задачей является восстановление распределений объектов в пространстве, что позволило бы перейти от анализа точечных примеров из выборки к анализу объектов в совокупности.

Широкий спектр современных моделей векторизации данных, разрабатываемых на основе нейронных сетей, в частности архитектуры *transformer* [33], в процессе своего построения и обучения не придавали специального смысла абсолютному значению векторов, что исходит из вида самой операции *attention* [33]. Вместе с этим и большое число современных текстовых эмбеддеров, такие как **BERT** [10] подобные, решают данную проблему принудительной нормировкой векторов на единичную сферу. Это приводит к оправданности использования **Directional statistics**, или же статистик на сфере, тем самым отбрасывая норму векторов. Для моделирования распределений в данном пространстве использовалась аппроксимация при помощи смеси **von Mises-Fisher** распределений, подробнее о чем в Appendix A

3.3 Нахождение компонент смесей

Для восстановления смесей используется EM алгоритм для смеси **von Mises-Fisher** распределений, описываемый в статье [15]. Для определения числа компонент и начального приближения использовался алгоритм **Mean-Shift**, ядром которого является косинусная метрика. Данный метод инициализации не требует задачи числа кластеров, способен находить "выбросы"в данных, устойчив в нахождении мод относительно размера выборки, что позволяет использовать его для определения число компонент смеси и принадлежности объектов найденным кластерам.

4 Задача разделимости распределений

Восстановление распределений классов T в вещественном пространстве эмбеддингов позволяет определить для любого вектора на $\mathbb{S}_n$ оценки плотностей вероятности его принадлежности конкретной компоненте каждого распределения. В свою очередь это делает возможным решение ряда задач на принятие/отвержение гипотез, получения оценок качества оператора-эмбеддера.

Оценка разделимости двух распределений является показателем, насколько качественно в принципе мы способны решить данные задачи (как точечное наблюдение позволяет судить о ее принадлежности тому или иному классу). Парную идентифицируемость будем определять индексом разделимости (или, как его называют в некоторых источниках, Баесовским индексом различимости), который имеет корни в классической задаче о простых гипотезах в теории **Signal Detection** [16] (где он известен как **discriminability index**).
Определим такой индекс через метрику $i(f_1, f_2)$ от функций плотностей распределений, действующих в одном вероятностном пространстве. Очевидно, что не каждая метрика способна выступать в качестве такой оценки. Для адекватности получаемых результатов, метрика должна учитывать разницу на всей сфере и отражать ее (например метрика Леви не подходит под данный параметр), быть аналитически вычислимой и устойчивой к размерности пространства и параметрам распределений. Классической метрикой является d' [16], развитие которой для случая распределений на поверхности сферы рассматриваются в данном исследовании. Далее предлагается аналитический вывод и способы использования двух метрик, основанных на Kullback-Leibler и χ^2 дивергенциях Appendix B

4.1 Метрика KL

Широко используемая оценка различия распределений, далее будем рассматривать индекс разделимости на основе данной дивергенции как $i_{KL} = \frac{KL(f_1\|f_2)+KL(f_2\|f_1)}{2}$. Такая метрика оценивает, насколько хорошо одно распределение приближает дифференциальный энтропийный параметр второго, задача, на которую изначально обучались векторизаторы. Конечная форма представима в виде Appendix B.1:

$$\begin{aligned} KL(f_1\|f_2) =& \eta \log\left(\frac{k_1}{k_2}\right) - \log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right) + \\ &+ r_\eta(k_1)\left\langle k_1\mu_1 - k_2\mu_2, \mu_1\right\rangle \end{aligned}$$

4.2 Метрика χ^2

Точно так же определим метрику как среднее от дивергенций в обоих направлениях $i_{\chi^2} = \frac{\chi^2(f_1\|f_2)+\chi^2(f_2\|f_1)}{2}$. Получаемая метрика так же, как и

i_{KL} не является ограниченной, но при этом обладает более экстремальным поведением на парах "близких"и "дальних"распределениях (стремление к нулю или к бесконечности выражено много сильнее, чем для метрики предыдущего пункта). Конечный вид представим как Appendix B.2:

$$\chi^2(f_1 \| f_2) = \frac{k_1^{2\eta} I_\eta(k_2) I_\eta(\overline{k})}{k_2^\eta \overline{k}^\eta I_\eta^2(k_1)} - 1$$

5 Вычисление статистик

Для восстановленной смеси распределений **von Mises-Fisher** параметры распределения будем находить как значения, доставляющие максимум для оценки правдоподобия выборки:

$$\mathscr{L}(\overline{X}, \mu, k) = \prod_{x \in \overline{X}} c_n(k) \exp\{\langle\, k\mu, x\rangle\}$$
$$\Longrightarrow \log \mathscr{L}(\overline{X}, \mu, k) = k \sum_{x \in \overline{X}} \mu^\intercal x + \log c_n(k) \to \max_{\mu^\intercal \mu = 1}$$

Из чего следует решение для параметров μ и k:

$$\mu = \frac{\sum_{x \in \overline{X}} x}{\| \sum_{x \in \overline{X}} x \|}$$
$$k = M_{n/2-1}^{-1} \left(\frac{\| \sum_{x \in \overline{X}} x \|}{n} = \overline{D} \right) = \left(\frac{I_{n/2}(\cdot)}{I_{n/2-1}(\cdot)} \right)^{-1} (\overline{D})$$

Задача устойчивого вычисления данных статистик и получения на их основе индексов разделимости подробнее рассматриваются в Appendix C и Appendix D

6 Получаемый алгоритм

Конечный вид предлагаемого алгоритма получения оценок на различимость классов рубрикатора представим как:

Algorithm 1 Обработка данных и вычисление индексов разделимости

Input: Обучающая выборка $X = \{x_1, x_2, \ldots, x_n\}$
Output: Индексы разделимости $I = \{i_1, i_2, \ldots, i_k\}$

Шаг 1: Векторизация обучающей выборки Преобразовать каждый объект x_i в выборке X в векторное представление v_i Сформировать коллекции объектов $V = [v_1, v_2, \ldots, v_T]$

Шаг 2: Восстановление смесей von Mises-Fisher Запустить метод **Mean-Shift**, получить начальное приближение и запустить EM алгоритм.

Шаг 3: Получение индексов разделимости Рассчитать меры разделимости между парами классов по каждой возможной паре компонент смеси Получить попарные индексы разделимости I, характеризующие качество различимости двух пар компонент смесей

return I

После выполнения данных операций, исследователь остается с матрицей, обхекты которых являются множество попарных расстояний между компонентами смесей разных классов рубрикатора. В качестве агрегирующей функции по таким наборам используется отношение **min** как наиболее подходящая в контексте конечной задачи исследования –нахождения самых близких друг к другу ценностей, разделяемых разметчиками.

7 Эксперименты

Разработанный инструмент статистического анализа позволяет формально моделировать вероятностную природу объектов в удобном для работы пространстве эмбеддингов, без необходимости учитывать их исходную форму, что достигается путем рассмотрения модели отображения объектов как черного ящика. Таким образом, становится возможным решать как строгие *прямые* задачи классификации наблюдаемых объектов на основе определения порождающего распределения и принятия или отвержения гипотез, так и *обратные* задачи оценки адекватности наблюдаемой выборки. Последний аспект особенно важен, так как включает в себя два ключевых вопроса, связанных с размеченными данными:

1. При условии корректного отображения объектов в пространство эмбеддингов, насколько хорошо и адекватно справились разметчики со своей работой.

2. При условии корректной разметки, насколько успешно эмбеддер создал адекватные представления объектов.

Каждая из полученных постановок задач основывается на определенных идеальных условиях, которые трудно полностью соблюсти на практике. Тем не менее, посредством интерпретации индексов разделимости распределений, рассматриваемых в исследовании, эти задачи могут быть решены на эмпирическом уровне. От инструмента ожидается устойчивость к большим размерностям пространства эмбеддингов и малым размерам обучающих выборок, соответствие значений индексов адекватным ожиданиям, наглядность в отображении свойств данных, а также способность решать задачи классификации объектов не хуже более простых прямолинейных методов.

7.1 Датасет

В исследовании использовался датасет над корпусом русских текстов, размеченных лабораторией Семантического анализа при университете МГУ. Далее будем отождествлять понятие **метка** и соответсвующий ей **класс**. Рассматривалась разметка ценностей в текстах по классификатору, состоящему из меток относящих ценность к соответствующим этническим или социальным группам, выражающих материальность/духовность ценности в различном проявлении, черты или характер проявления данной

ценности. Разметка соответствует задаче multilabel-classification, где каждый фрагмент помечен набором меток, определяющих, какая ценность в конкретном фрагменте содержится, какое отношение к ней сложилось у автора, какое воздействие приведение данной ценности должно было произвести на читающего. Такой функционал отчасти выражается в
существовании 4 специальных меток тональности фрагмента: *отрицательная*, *нейтральная*, *положительная*, *конфликт тональности*.

Данная задача декомпозирована на задачу multilabel-classification, где каждый объект, помеченный набором меток, представлялся как множество одинаковых объектов, каждый из которых относится к одному
соответствующему классу по изначальной группе.

7.2 Постановка задачи

Пусть определены множества документов $\mathscr{D} = \{d_1, ..., d_m\}$, множество классов $T = \{t_1, ..., t_\tau\}$, всего фрагментов F, разницей в работе разметчиков можно пренебречь, считая их равными, тем самым не рассматривая отдельно взятого разметчика как случайное воздействие. Тогда будем считать токенами в тексте (неделимыми по смыслу единицами) слова. Разметчик, смотря на слово в контексте документа, порождает дискретное распределение на множестве классов T

$$p(t_k|token_i, d_j), t_i \in T, token_i \in d_j, d_j \in \mathscr{D}$$

Откуда случайным образом выбирает метку и ставит ее в соответствие слову $token_i$. И оптимизационная задача становится задачей максимизации оценки правдоподобия:

$$L_{prob} = \prod_i p(t_i|token_i, d_{j_i}) \longrightarrow \max_{t_i \in T}$$

Пусть существует оператор, переводящий фрагменты текста осмысленно в пространство, интерпретируемое компьютером. Действуя в изначальных предположениях об этом операторе, его способности схожие по смыслу фрагменты ставить ближе, чем фрагменты с противоположным или контрастирующим содержанием (гипотеза компактности в пространстве смыслов озвученная во вступлении), будем приближать данный оператор работой **LLM**. В экспериментах использовался **ruBERT-base-cased-sentence**, обученный на задаче предсказания фрагментов текста.

7.3 Анализ распределений классов

Рассматривалась выборка из 84 классов, разметка по которым присутствует в датасете.

1. Для каждого класса по отдельности были найдены все области сгущения представителей с помощью алгоритма кластеризации **MeanShift**, каждая из которых сопоставлялась отдельному распределению.

2. Применялся алгоритм кластеризации EM, на основе работы которого находились оптимальные параметры μ и κ.

3. Для каждой пары компонент разных смесей находились индексы i_{KL} и i_{χ^2}, после чего брался **min** для получения оценки разделимости смесей между собой.

В результате проведения данной операции получено 141 компонента, из которых для каждого отдельного класса соответствует не более 2. Тем самым каждый класс соответствует унимодальному или бимодальному распределению, а потому в предположениях об идейной содержательности такого представления, каждый класс несет в себе одну общую, либо две достаточно разных идеи. Теперь возможен анализ, какие классы между собой разделяют наиболее близкие и наиболее дальние идеи. Для этого будем брать наименьшее расстояние между составными распределениями двух классов. Результаты представлены ниже:

Группа 1: Самые близкие друг к другу классы

Таблица 1: Метрика i_{χ^2}

Класс 1	Класс 2	i
Достоинство	Культура и искусство	7.45e-9
Еда	Важность общ. мнения	3.17e-8
Достоинство	Честность	3.43e-8
Достоинство	Историческая память	5.87e-8
Гордость	Потворство желаниям	7.478e-8
Любовь	Культура и искусство	1.21e-7
Этничность	Удовольствие жизнью	6.27e-7
Воспитание	Познание	1.19e-6
Образование	Благочестие	1.72e-6
Патриотизм	Природа	2.44e-6

Таблица 2: Метрика i_{KL}

Класс 1	Класс 2	i
Мат. ценности	Нейтральный тон	36.86
Мор. ценности	Чувство принадлежности	78.77
Смелость	Положительный тон	115.85
Патриотизм	Уважение традиций	122.34
Полит. ценности	Права и свободы	126.34
Заботливость	Положительный тон	130.55
Нац. безопасность	Выборность власти	133.64
Жизнь	Чувство принадлежности	134.82
Безопасность	Конфликт тональности	135.90
Патриотизм	Этничность	176.16

Группа 2: Самые удаленные друг к другу классы

Таблица 3: Метрика i_{χ^2}

Класс 1	Класс 2	i
Язык	Справедливость	6.24e16
Счастье	Твердая воля	1.40e15
Справедливость	Творчество	7.36e14
Заботливость	Нейтральный тон	2.88e13
Гуманизм	Смелость	1.69e13
Культура	Здоровье	5.14e11
Творчество	Смелость	2.46e11
Язык	Счастье	1.95e11
Творчество	Заботливость	5.44e10
Счастье	Жилище	2.50e10

Таблица 4: Метрика i_{KL}

Класс 1	Класс 2	i
Историческая память	Впечатления от жизни	7465.42
Смелость	Гуманизм	6651.31
Этничность	Мат. ценности	6433.21
Природа	Образование	6379.33
Талант	Чувство юмора	6235.82
Красота	Впечатления от жизни	6190.90
Власть	Жизнь	6186.52
Творчество	Верность	6174.12
Язык	Честность	6165.45
Еда	Твердая воля	5818.10

Полученные результаты для обеих метрик по нахождению 10 наиболее близких и отдаленных друг от друга классов хорошо согласуются с эмпирическим опытом и ожиданиями разметчиков/людей. Тем не менее в зависимости от выбора метрики состав таблицы меняется. Так, метрика на основе KL-дивергенции смогла определить близкими между собой не только смысловые классы, но и классы, определяющие тональности фрагментов. Вместе с этим и разброс значений метрик рознится: i_{χ^2} принимает значения от 10^{-9} до 10^{16}, когда для i_{KL} разброс представлен разницей в два порядка. Важно отметить, что в каждой из таблиц присутствует доля повторяющихся несколько раз тегов. Такое поведение объясняется высокой схожестью фрагментов, представляющих соответственные классы, в контексте данного корпуса текстов.

7.4 Работа с результатами

Предлагаемый алгоритм ставит первостепенной задачей указать на возможные проблемы в организации рубрикатора:

- Разметчики (эксперты) не способны увидеть разницы между двумя классами в действительных данных, что противоречит первоначальной гипотезе, что данные классы сильно различаются по своей семантике – на это указывает сравнительно малое расстояние между распределениями семантически различных классов.
- Наборот, классы, что близки по своей направленности, воспринимаются раметчиками как взаимоисключающие классы.

Дальнейшая работа с датасетом культурный код позволила выявить ряд особенностей в данных, связанных с выдаваемой инструкцией для разметчиков, а так же переработать иерархию классов в рубрикаторе, основываясь на общении с лингвистами.

8 Дальнейшая работа

В данный момент основной проблемой метода является ненормированность метрик разделимости, что усложняет анализ пар классов в отрыве от всего множества оценок на $T \times T$. Вместе с этим не решен вопрос получения честных оценок на разделимость смесей, что является трудно-вычислимой задачей при простейших подходах, из-за чего не рассматривались в данной работе.

9 Выводы

В данном исследовании мы предложили новый метод мультикритериальной оценки качества рубрикатора, позволяющий проводить анализ рубрикатора напрямую с экспертами-составителями рубриктора. Предложенный метод решает проблему скупости информативности более классических методов, в отличие от которых использует статистические методы для агрегации выборки и высчитыванию метрик качеств на более богатом семействе объектов – смесях распределений. Получаемые значения позволяют во время активной фазы разработки задания для разметчиков или же на первых стадиях их работы вносить коррективы в сформированный рубрикатор.

Список литературы

[1] Advanced signal processing handbook: Theory and implementation for radar, sonar, and medical imaging real time systems, December 2000.

[2] IA Bolshakov and A Gelbukh. Classification of collocations in databases by meaning of combined words. *AUTOMATIC DOCUMENTATION AND MATHEMATICAL LINGUISTICS TRANSLATIONS OF SELECTED ARTICLES FROM NAUCHNO-TEKHNICHESKAIA INFORMATSIIA*, 34(3):64–74, 2000.

[3] Francois Boutin and Mountaz Hascoët. Cluster validity indices for graph partitioning. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 376–381. IEEE, 2004.

[4] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[5] Alan D Chave. A note about gaussian statistics on a sphere. *Geophysical Journal International*, 203(2):893–895, 2015.

[6] Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825, 2024.

[7] Zi Li Chen. Research and application of clustering algorithm for text big data. *Computational Intelligence and Neuroscience*, 2022(1):7042778, 2022.

[8] Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1197–1203. IEEE, 1999.

[9] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[10] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[12] Akhmedov Farkhod, Akmalbek Abdusalomov, Fazliddin Makhmudov, and Young Im Cho. Lda-based topic modeling sentiment analysis using topic/document/sentence (tds) model. *Applied Sciences*, 11(23):11091, 2021.

[13] Inna E Gendlina. The russian rubricator: A unified system of classificatory indexing languages. *KO KNOWLEDGE ORGANIZATION*, 19(3):126–130, 1992.

[14] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

[15] Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *International Conference on Machine Learning*, pages 154–162. PMLR, 2014.

[16] Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. *Detection theory: A user's guide*. Routledge, 2021.

[17] Jinwen He, Yujia Gong, Zijin Lin, Yue Zhao, Kai Chen, et al. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10218–10230, 2024.

[18] Swapnil Hingmire, Sandeep Chougule, Girish K Palshikar, and Sutanu Chakraborti. Document classification by topic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 877–880, 2013.

[19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[20] Wei Lu and Jagath C Rajapakse. Approach and applications of constrained ica. *IEEE transactions on neural networks*, 16(1):203–212, 2005.

[21] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[22] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[23] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[24] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 24(12):1650–1654, 2002.

[25] David A McConnell, David N Steer, and Kathie D Owens. Assessment and active learning strategies for introductory geology courses. *Journal of Geoscience Education*, 51(2):205–216, 2003.

[26] MES Mendes and Lionel Sacks. Dynamic knowledge representation for e-learning applications. In *Enhancing the Power of the Internet*, pages 259–282. Springer, 2004.

[27] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[28] Arturo Montejo Ráez et al. Automatic text categorization of documents in the high energy physics domain. 2005.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[30] Barbara Sakitt. Indices of discriminability. *Nature*, 241(5385):133–134, 1973.

[31] Nayani Sateesh, Kuljeet Kaur, M Lakshminarayana, Vipul Vekariya, Harshal Patil, and Ramya Maranan. Development of a gui for automated classification of scientific journal articles using clustering. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE, 2024.

[32] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of i s (x). *Computational Statistics*, 27:177–190, 2012.

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[34] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.

[35] Dejian Yu and Bo Xiang. Discovering topics and trends in the field of artificial intelligence: Using lda topic modeling. *Expert systems with applications*, 225:120114, 2023.

[36] Oren Eli Zamir. *Clustering web documents: a phrase-based method for grouping search engine results*. University of Washington, 1999.

[37] Weiwen Zou and Pong C Yuen. Discriminability and reliability indexes: two new measures to enhance multi-image face recognition. *Pattern Recognition*, 43(10):3483–3493, 2010.

[38] Руслан Айдарович Гилязев and Денис Юрьевич Турдаков. Активное обучение и краудсорсинг: обзор методов оптимизации разметки данных. *Труды Института системного программирования РАН*, 30(2):215–250, 2018.

[39] М Дли, О Булыгина, П Козлов, and В Борисов. *Rubrication of text documents based on fuzzy difference relations*. Litres, 2022.

[40] Татьяна Сергеевна Ильина and Татьяна Сергеевна Ившина. Выявление духовных ценностей в текстах семейных преданий. *Балтийский гуманитарный журнал*, 7(3 (24)):41–45, 2018.

[41] ЯБ Калачев and АН Сибирмовская. Создание кластерного пространства текстовых документов в базе данных. *Новые информационные технологии в автоматизированных системах*, (13):130–133, 2010.

[42] ОВ Пескова. Методы автоматической классификации текстовых электронных документов. *Научно-техническая информация. Серия 2: Информационные процессы и системы*, (3):13–20, 2006.

[43] Ольга Вадимовна Пескова. Автоматическое формирование рубрикатора полнотекстовых документов. In *Тр. Десятой Всерос. науч. конф.«Электронные библиотеки: перспективные методы и технологии, электронные колекции»(RCDL'2008).—Дубна*, pages 139–148, 2008.

Appendices

A Выбор семейства распределений

Основной задачей является нахождение качественной аналитической аппроксимациии распределений для классов в пространстве векторизованных представлений. Однако использование простейших параметрических семейств не применимо в данном случае не применимо в виду отсутствия у нас априорных знаний об приближаемых объектах. Решение данной задачи предлагает упоминаемая в 3 главе [1] теорема «**Approximation Theorem**», обозначающая возможность аппроксимации любого многомерного распределения Гауссовской смесью, что широко применяется на практике. Однако область определения распределения Гауссовских смесей является все пространство $\mathbb{R}^n$, что не подходит для задачи на сферической поверхности. Используя условие периодичности по 2π и нормировки на 1 функция плотности гауссовской величины для $\mathbb{R}^2$ примет вид [5]:

$$\begin{cases} WN(\theta;\mu,\sigma) \sim \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\sum_{k=-\infty}^{\infty} \exp\left[\frac{-(\theta-\mu+2\pi k)^2}{2\sigma^2}\right] \\ f_{WN}(\theta;\mu,\sigma) = \dfrac{1}{2\pi}\vartheta\left(\dfrac{\theta-\mu}{2\pi}, \dfrac{i\sigma^2}{2\pi}\right) \\ \vartheta(\theta,\tau) = \displaystyle\sum_{n=-\infty}^{\infty} (w^2)^n q^{n^2}, \text{ где } w \equiv e^{i\pi\theta} \\ \vartheta(\theta,\tau) - \text{Тэта функция Якоби} \end{cases}$$

Расчет подобных формул представляется неэффективным в контексте высоко нагруженных задач даже для простейшего случая размерности $n = 2$. При этом положение лишь усугубляется при переходе к $n > 2$ мерному случаю [5]. Определим условия для нового кандидата на роль распределения, по которому будет восстанавливаться смесь распределений:

1. Распределение должно быть унимодальным с центром в моде.
2. При удалении от центра плотность должна стремительно убывать к нулю.
3. Распределение подобно Гауссовскому должно обладать параметром кучности.

Таковым распределением предлагается рассматривать известное распределение из циркулярной статистики, приближающее Гауссовское на сфере — Распределение **von Mises-Fisher**.

Определим поверхность сферы как пространство элементарных исходов. Важнейшей особенностью является конечность меры поверхности. Тогда элементарным приращением $dw = dx_1 \times ... \times dx_{d-1} = dx$, что через якобиан переходит в углы по формуле: $dw = \left(\prod_{i=2}^{p-2} \sin^{p-i}\theta_{i-1}\right) d\theta$. А само распределение будет выводиться, как зависящее от параметра кучности k и центра распределения μ:

$$\begin{cases} p(x|\mu,k) = c_n(k)e^{k\langle\mu,x\rangle} \\ c_n(k) = \dfrac{k^{n/2-1}}{(2\pi)^{n/2}I_{n/2-1}(k)} \\ I_n(k) = \left(\dfrac{k}{2}\right)^n \displaystyle\sum_{t\geq 0} \frac{1}{\Gamma(n+t+1)t!}\left(\frac{k}{2}\right)^{2t} \end{cases}$$

- k является параметром кучности распределения, и чем он ближе к нулю, тем более равномерно распределение.
- μ является центром распределения.
- $c_n(k)$ есть нормировочная константа для распределения.
- С ростом размерности n резко увеличиваются проблемы с высчитываем параметров из-за float-point вычислений

Для такого распределения первый момент $\mathbb{E}x = M_{n/2-1}(k)\mu$, где $M_{n/2-1}(k) = \frac{I_{n/2}(k)}{I_{n/2-1}(k)}$ есть функция от параметра k и размерности пространства. В свою очередь данная функция сложна в вычислении для больших размерностей n, что решается далее в статье.

B Вывод формул для индексов разделимости

Статистика $d'(\cdot,\cdot)$ является наиболее классической оценкой разделимости распределений в теории **Signal Detection**, где была выведена из для задачи о разделимости двух нормальных распределений соответствующих гипотезе H_0 и H_1. Для двух гипотез и выбранного порога, статистика $d' = z(P(\text{True positive})) - z(P(\text{False positive}))$, $d' \in (-\infty, +\infty)$, где функция $z(\cdot)$ есть обратная к нормальной функции распределения, и разность вероятностей правильно определить класс H_1 при попарном сравнении с классом H_0 для заданного критерия, из-за чего данная метрика напрямую не применима к введенным ранее распределениям. Попытка же адаптировать d' для распределения **von Mises-Fisher** так же является сложной задачей, ведь подразумевает аналитический вывод кумулятивной функции распределения на сфере, что вынуждает строить иные индексы, способными имитировать работу d'.

B.1 Метрика Kullback-Leibler

Рассмотрим, чему в общем случае равна дивергенция по поверхности сферы:

$$KL(f_1\|f_2) = \int_{\mathbb{S}_{p-1}} f_1(x|k_1,\mu_1)\log\left(\frac{f_1(x|k_1,\mu_1)}{f_2(x|k_2,\mu_2)}\right)dx = \int_{\mathbb{S}_{p-1}}\left(\langle k_1\mu_1, x\rangle - \langle k_2\mu_2, x\rangle + \log\left(\frac{c_n(k_1)}{c_n(k_2)}\right)\right) f_1(x|k_1,\mu_1)dx$$

Тогда интегрирование обеих частей дает систему:

$$\begin{cases} \displaystyle\int_{\mathbb{S}_{p-1}} \langle k_1\mu_1 - k_2\mu_2, x\rangle\, f_1(x|k_1,\mu_1)dx = \langle k_1\mu_1 - k_2\mu_2, \mathbb{E}x_1\rangle \\ \displaystyle\int_{\mathbb{S}_{p-1}} \log\left(\frac{c_n(k_1)}{c_n(k_2)}\right) f_1(x|k_1,\mu_1)dx = \log\left(\frac{c_n(k_1)}{c_n(k_2)}\right) \end{cases}$$

Тогда получим выражение дивергенции ($\eta = n/2 - 1$):

$$KL(f_1\|f_2) = \eta\log\left(\frac{k_1}{k_2}\right) - \log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right) + r_\eta(k_1)\,\langle k_1\mu_1 - k_2\mu_2, \mu_1\rangle$$

B.2 Метрика χ^2

$$\chi^2(f_1\|f_2) = \int_{\mathbb{S}_{p-1}} \frac{\left(f_1(x|k_1,\mu_1) - f_2(x|k_2,\mu_2)\right)^2}{f_2(x|k_2,\mu_2)}dx$$

Раскрывая скобки, упростим интеграл в условиях независимости распределений. Обозначим $\eta = 2/n - 1$:

$$\chi^2(f_1\|f_2) = \int_{\mathbb{S}_{p-1}} \frac{\left(f_1(x|k_1,\mu_1)\right)^2}{f_2(x|k_2,\mu_2)}dx - 1 \Rightarrow \frac{\left(f_1(x|k_1,\mu_1)\right)^2}{f_2(x|k_2,\mu_2)} = \exp\left\{\langle 2k_1\mu_1 - k_2\mu_2, x\rangle\right\}\frac{c_\eta^2(k_1)}{c_\eta(k_2)}$$

Заметим, что получаемое выражение повторяет запись для распределения **von Mises-Fisher** со специальными параметрами: $\overline{\mu} = \frac{(2k_1\mu_1 - k_2\mu_2)}{\|2k_1\mu_1 - k_2\mu_2\|_{\ell^2}}$; $\overline{k} = \|2k_1\mu_1 - k_2\mu_2\|_{\ell^2} \Longrightarrow$

$$\chi^2(f_1\|f_2) = \frac{c_n^2(k_1)}{c_n(k_2)c_n(\overline{k})} - 1 = \frac{k_1^{2\eta} I_\eta(k_2) I_\eta(\overline{k})}{k_2^\eta \overline{k}^\eta I_\eta^2(k_1)} - 1$$

C Устойчивость вычислений параметров

Для параметров:

$$\mu = \frac{\sum_{x\in\overline{X}} x}{\|\sum_{x\in\overline{X}} x\|} \quad k = M_{n/2-1}^{-1}\left(\frac{\|\sum_{x\in\overline{X}} x\|}{n} = \overline{D}\right) = \left(\frac{I_{n/2}(\cdot)}{I_{n/2-1}(\cdot)}\right)^{-1}(\overline{D})$$

Нахождение устойчиво вычислимой обратной функции для M в общем случае не представляется возможным, а потому используются приблизительные оценки. Данная статья использует простейший способ, не включающий в себя использование функций Бесселя, что тем не менее с ростом размерности пространства и выборки асимптотически стремится к действительному решению:

$$\mu = \frac{\sum_{x\in\overline{X}} x}{n\overline{D}} \quad k = \frac{\overline{D}(n - \overline{D}^2)}{1 - \overline{D}^2}$$

Полученные параметры μ и k позволяют без явного построения получить по выведенным формулам индексы разделимости, и как итог, проанализировать распределения в пространстве эмбеддингов эффективным образом. Однако для каждого индекса i так или иначе требует вычисления функций Бесселя I

высоких порядков, а так же произведений и делений экстремально малых или экстремально больших по модулю чисел. Далее рассматривается численно-устойчивый итерационный метод расчета I [32]

C.1 Вычисление функции Бесселя

Модернизированная первая функция бесселя представима в виде степенного ряда:

$$I_n(k) = \left(\frac{k}{2}\right)^n \sum_{t\geq 0} \frac{1}{\Gamma(n+t+1)t!}\left(\frac{k}{2}\right)^{2t}$$

Точное получение значения такой функции затруднительно с точки зрения достижения вычислительной устойчивости, так как в степенном ряду для каждого члена приходится высчитывать значение дроби (на практике) со стремящимися к бесконечности знаменателем и числительным. При этом подсчет точного значения гамма-функции может быть вычислительно неэффективным при больших n. В данном случае мы стремимся упростить выражение, разбить его на части, чтобы избежать неустойчивых выражений при вычислении индексов i [32]. Для этого, используя свойство Гамма-функции $\Gamma(x+1) = x\cdot\Gamma(x)$ и ее аппроксимацию Стирлинга, упростим выражение:

$$\begin{cases} I_n(k) = \dfrac{k^n}{2^n\Gamma(n)} \displaystyle\sum_{t\geq 0} \frac{1}{t(t+1)(t+2)...(t+n)t!}\left(\frac{k}{2}\right)^{2t} = \frac{k^n}{2^n\Gamma(n)}\cdot Tail(n,k) \\ Tail(n,k) = \displaystyle\sum_{t\geq 0} \frac{1}{t(t+1)(t+2)...(t+n)t!}\left(\frac{k}{2}\right)^{2t} \\ \Gamma(n) = \left(\dfrac{n}{e}\right)^n \sqrt{\dfrac{2\pi}{n}}\left(1+\dfrac{1}{12n}+\dfrac{1}{288n^2}+\overline{o}\left(\dfrac{1}{n^2}\right)\right) \end{cases}$$

Тогда высчитывание функции $I_n(k)$ производится итерационным способом. Член степенного ряда получается из предыдущего умножением на $\frac{k^2/4}{(t+1)(n+t+1)}$. Тогда получаем алгоритм подсчета, представленный ниже:

Algorithm 2 *Алгоитм вычисления первой модифицированной функции Бесселя* $I_n(ka)$

Result: Input n $\gg 1$, $k \geq 0$
$T \leftarrow 1.0$
$A' \leftarrow \left(\frac{ke}{2n}\right)^n \cdot \sqrt{\frac{n}{2\pi}}$
$A \leftarrow A'/\left(1+\frac{1}{12n}+\frac{1}{288n^2}\right)$
$Tail \leftarrow 0$; $t \leftarrow 1$ **while** $\Delta T > \varepsilon(A)$ **do**
end
$T \leftarrow T \cdot \frac{k^2/4}{(t+1)(n+t+1)}$
$Tail \leftarrow Tail + T$
$t \leftarrow t+1$

return $A \cdot Tail$

D Стабильное вычисление индексов разделимости

Всюду далее будет использоваться двухместная функция $Tail(\cdot,\cdot)$, определенная в Appendix C.1

D.1 Вычисление KL дивергенции

Вернемся к записи $KL(f_1\|f_2)$ и применим представление для $I_n(k)$, взяв $\eta = n/2-1$

$$KL(f_1\|f_2) = \eta\log\left(\frac{k_1}{k_2}\right) - \log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right) + r_\eta(k_1)\left\langle k_1\mu_1 - k_2\mu_2, \mu_1\right\rangle$$

Рассмотрим второй член выражения:

$$\log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right) = \log\left(\frac{\left(\frac{k_1 e}{2\eta}\right)^\eta \frac{\sqrt{\eta/2\pi}}{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)} Tail(\eta,k_1)}{\left(\frac{k_2 e}{2\eta}\right)^\eta \frac{\sqrt{\eta/2\pi}}{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)} Tail(\eta,k_2)}\right) = \eta\log\left(\frac{k1}{k2}\right) + \log\left(\frac{Tail(\eta,k_1)}{Tail(\eta,k_2)}\right)$$

Рассмотрим представление $r_\eta(k_1)$ и выведем ее ассимптотическое поведение при $\eta \to \infty$:

$$r_\eta(k_1) = \frac{I_{\eta+1}(k_1)}{I_\eta(k_1)} = \left(\frac{k_1 e}{2(\eta+1)}\right)^{\eta+1} \cdot \left(\frac{2\eta}{k_1 e}\right)^{\eta} \cdot \sqrt{\frac{\eta+1}{\eta}} \cdot \frac{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)}{\left(1+\frac{1}{12(\eta+1)}+\frac{1}{288(\eta+1)^2}\right)} \cdot \frac{Tail(\eta+1,k_1)}{Tail(\eta,k_1)} =$$

$$= \frac{k_1 e}{2(\eta+1)} \cdot \left(1-\frac{1}{\eta}\right)^{\eta} \cdot \sqrt{1+\frac{1}{\eta}} \cdot \frac{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)}{\left(1+\frac{1}{12(\eta+1)}+\frac{1}{288(\eta+1)^2}\right)} \cdot \frac{Tail(\eta+1,k_1)}{Tail(\eta,k_1)} \xrightarrow[\eta\to\infty]{} \frac{k_1}{2(\eta+1)} \frac{Tail(\eta+1,k_1)}{Tail(\eta,k_1)}$$

Тогда в предположении $n \gg 1$ полная формула получения $KL(f_1\|f_2)$ будет иметь вид:

$$KL(f_1\|f_2) \cong -log\left(\frac{Tail(\eta,k_1)}{Tail(\eta,k_2)}\right) + \frac{k_1}{2(\eta+1)} \frac{Tail(\eta+1,k_1)}{Tail(\eta,k_1)} \langle k_1\mu_1 - k_2\mu_2, \mu_1\rangle$$

$$\Longrightarrow i_{KL} = \frac{k_1}{2(\eta+1)} \frac{Tail(\eta+1,k_1)}{Tail(\eta,k_1)} \langle k_1\mu_1 - k_2\mu_2, \mu_1\rangle + \frac{k_2}{2(\eta+1)} \frac{Tail(\eta+1,k_2)}{Tail(\eta,k_2)} \langle k_2\mu_2 - k_1\mu_1, \mu_2\rangle$$

D.2 Вычисление χ^2 дивергенции

Так же рассмотрим запись χ^2 дивергенции, и упростим, используя формулы для $I_n(k)$

$$\chi^2(f_1\|f_2) = \frac{c_n^2(k_1)}{c_n(k_2)c_n(\overline{k})} - 1 = \frac{k_1^{2\eta} I_\eta(k_2) I_\eta(\overline{k})}{k_2^\eta \overline{k}^\eta I_\eta^2(k_1)} - 1$$

Тогда подставим формулу $I_n(k)$:

$$\chi^2(f_1\|f_2) = \frac{k_1^{2\eta}}{k_2^\eta \overline{k}^\eta} \cdot \frac{\left(\frac{k_2 e}{2\eta}\right)^\eta \left(\frac{\overline{k} e}{2\eta}\right)^\eta Tail(k_2,\eta) Tail(\overline{k},\eta)}{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)^2} \cdot \frac{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)^2}{\left(\frac{k_1 e}{2\eta}\right)^{2\eta} Tail^2(k_1,\eta)} = \frac{Tail(k_2,\eta)Tail(\overline{k},\eta)}{Tail^2(k_1,\eta)}$$

И соответственный индекс разделимости распределений i_{χ^2} примет вид:

$$\Longrightarrow i_{\chi^2} = \frac{Tail(k_2,\eta)Tail(\overline{k},\eta)}{Tail^2(k_1,\eta)} + \frac{Tail(k_1,\eta)Tail(\overline{k},\eta)}{Tail^2(k_2,\eta)}$$