Московский Физико-Технический Институт
(Государственный Университет)

Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 174 ГРУППЫ

«Применение коллаборативной фильтрации в задаче выделения селекторов»

Выполнил:
студент 4 курса 174 группы
Шинкевич Михаил Игоревич

Научный руководитель:
д.ф-м.н.
Воронцов Константин Вячеславович

Москва, 2015

Содержание

Аннотация

Рассмотрена задача персонализации рекомендаций. Используется коллаборативная фильтрация и FLAME кластеризация. Предложен метод выявления схожести вкусов пользователей на основе коллаборативный фильтрации без использования соц-демографической информации. Работа проиллюстрирована на сервисе А/Б-тестирования фотографий.

Ключевые слова: FLAME кластеризация, коллаборативная фильтрация.

1 Введение

На сегодняшний день задача выявления вкусов пользователей встречается в различных прикладных областях: построение разнообразных рекомендательных систем, таргетированная реклама и др. Помимо этого, в условиях современной реальности остро стоит вопрос приватности, поэтому необходимо уметь выявлять вкусы пользователей без использования их соц-демографической информации.

Целью данной работы является построение алгоритма персонализации рекомендаций пользователей на основе их вкусов, позволяющего улучшить качество онлайн-сервиса А/Б-тестирования фотографий: ускорить получение результата пользователем. При построении алгоритма использовались коллаборативная фильтрация и FLAME-кластеризация.

А/Б-тестирование (также говорят "сплит тестирование") заключается в сравнении двух версий (А и Б) одного объекта. Побеждает та версия, которая на одинаковых пользователях дает лучшую конверсию.

Будем называть мнением человека на данном А/Б-тестировании вариант(А или Б), который данный пользователь выбрал, а правильным результатом — среднее мнение всех пользователей сервиса. Среднее мнение принимает значения из множества натуральных чисел от 0 до 100 и выражено в процентах.

Задача заключается в определении среднего мнения на данном А/Б-тестировании при участии в тестировании наименьшего числа пользователей. Рассматриваемая задача является задачей кластеризации пользователей на основе их вкусов [1], [2]. Задачи подобного типа часто возникают в интернет проектах. Для их решения используют коллаборативную фильтрацию [3], [4], [5]. К сожалению, в условиях сегодняшних реалий, вкусы пользователей в коллаборативный фильтрации вычисляются косвенно, через схожесть объектов (фильмов, книг), т.к. каждый человек может посмотреть лишь небольшое количество фильмов. В данной же работе коллаборативная фильтрация используется для прямого, а не косвенного, сравнения вкусов пользователей, что обусловленно спецификой задачи. Функция похожести пользователей составляется на основе экспертных данных о факторах, демонстрирующих вкус пользователя [6], [7]. На основе информации о схожести вкусов пользователей выполняется их кластеризация. Для кластеризация использовался FLAME алгоритм

[8], [9], базирующийся на k-ближайших соседях. Для кластеризации выбран именно данный алгоритм в связи с тем, что он работает быстро(субквадратичное время) для большой базы пользователей и дает хороший результат, что было показано в исследованиях, проведенных командами Twitter, Snapchat, Tinder и др. [10], [11].

2 Основные гипотезы

Создавая А/Б-тестирование, пользователь хочет получить среднее мнение. Однако, возможна ситуация, когда пользователи, участвующие в данном А/Б-тестировании, принадлежат узкой группе пользователей с одинаковым вкусом и результат их тестирования сильно отличается от среднего мнения. В условиях мобильного сервиса целью пользователя является получение среднего мнения максимально быстро, поэтому целесообразным является кластеризация пользователей в группы с одинаковыми вкусами. Тогда в А/Б-тестировании будут принимать участие пользователи из разных кластеров..

3 Постановка задачи

Дано:

- Множество троек $T = \{u, c, r\}$ —

  {id пользователя, id А/Б-тестирования, выбор пользователя из А и Б}

- Функция схожести вкусов пользователей:

$$sim(u_i, u_j) = \alpha \frac{q_s^{ij}}{q_a^{ij}} + \beta \frac{w_s^{ij}}{q_a^{ij}} + \gamma \frac{s_b^{ij}}{s_a^{ij}},$$

  q_s^{ij} - количество голосований, на которых оба пользователя i и j сделали одинаковый выбор,

  q_a^{ij} - суммарное количество голосований, в которых принимали участие оба пользователя i и j,

  w_s^{ij} - взвешенное количество голосований, на которых оба пользователя i и j сделали одинаковый выбор,

  s_b^{ij} - количество голосований, на которых оба пользователя i и j отказались голосовать,

  s_a^{ij} - суммарное количество голосований, на которых отказался голосовать хотя бы один из пользователей i и j.

Требуется **найти** параметры α, β и γ, при которых доставляется минимум функции ошибки:

$$E(U) = \frac{1}{n} \sum_{i=1}^{n} [\frac{1}{m} \sum_{j=1}^{m} (\frac{out_{ij}}{all_i})],$$

out_{ij} - число пользователей из кластера i, сделавших в А/Б-тестировании j выбор не такой, как сделало большинство пользователей данного кластера,

all_i - число пользователей кластера i,

m - размер(количество А/Б-тестирований) тестовой выборки,

n - количество кластеров,

U - множество всех пользователей.

4 Описание FLAME алгоритма

Оптимизационная задача FLAME

Необходимо минимизировать функцию:

$$E(\{\boldsymbol{p}\}) = \sum_{\boldsymbol{x}\in\boldsymbol{X}} \left\| \boldsymbol{p}(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{N}}(\boldsymbol{x})} w_{\boldsymbol{xy}}\boldsymbol{p}(\boldsymbol{y}) \right\|^2$$

$\boldsymbol{X}$ - множество объектов всех 3 типов,

$\boldsymbol{p}(\boldsymbol{x})$ - вектор мягкого членства объекта $\boldsymbol{x}$,

$\mathcal{N}(x)$ - множество ближайших соседей $\boldsymbol{x}$,

$w_{\boldsymbol{xy}}$ - коэффициенты, отражающие близость соседа, $\sum\limits_{\boldsymbol{y}\in\boldsymbol{\mathcal{N}}(\boldsymbol{x})} w_{\boldsymbol{xy}} = 1$.

Функция ошибки может быть минимизирована решением следующих линейных уравнений:

$$p_k(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{N}}(\boldsymbol{x})} w_{\boldsymbol{xy}}p_k(\boldsymbol{y}) = 0, \quad \forall \boldsymbol{x} \in \boldsymbol{X}, \quad k = 1, ..., M$$

где M - число кластеров.

Следующая итеративная процедура может быть использована для решения этих линейных уравнений:

$$\boldsymbol{p}^{t+1}(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{N}}(\boldsymbol{x})} w_{\boldsymbol{xy}}\boldsymbol{p}^t(\boldsymbol{y})$$

Таким образом, FLAME алгоритм состоит из 3 этапов:

1) Выделение структурной информации данных

1. Построить граф соседства соединяющий каждый объект с его k-ближайшими соседями.
2. Оценить плотность каждого объекта на основе близости к своим соседям

$$\rho_j = \frac{max(\frac{1}{k}\sum_{i=1}^{k} d_i)}{\frac{1}{k}\sum_{i=1}^{k} d_i}$$

3. Объекты принадлежат одному из 3 классов:

a) Центр кластера: объекты с плотностью большей, чем все его соседи

b) Выбросы: объекты с плотностью меньшей, чем у всех соседей и ниже, чем задний порог

c) Все остальные объекты

2) Вычисление доли принадлежности каждому кластеру:

1. Инициализация членства

a) Каждому центру кластера присваивается полное членство своего кластера

b) Все выбросы присваиваются кластеру выбросов

c) Все остальные объекты получают равное членство каждого кластера

2. Итеративное обновление членства объектов на основе линейной комбинации членства своих соседей.

$$s_j = \sum_{i=1}^{k} (w_i s_i), \quad w_j = \frac{\frac{1}{d_j}}{\sum_{i=1}^{k} \frac{1}{d_i}}$$

3) Присвоение объектов кластер, в котором он имеет наибольшее членство

5 Описание решения

Для проведения эксперимента были взяты реальные данные онлайн-сервиса А/Б-тестирования фотографий. Данные представляют собой матрицу пользователь–А/Б-тестирование, в ячейках которой находится выбор(из А и Б) данного пользователя на данном тестировании. Количество пользователь было взято $|U| = 1000$, количество А/Б-тестирований, в которых все данные пользователи принимали участие $|C| = 10000$. Данная выборка(множество А/Б-тестирований) была разбита на обучающую и тестовую выборки в отношении 7 к 3. Обучение проходило следующим образом: сначала фиксировались коэффициенты значимости α, β и γ. Используя полученную функцию похожести вычисляла похожесть каждой пары пользователей.

$$sim(u_i, u_j) = \alpha \frac{q_s^{ij}}{q_a^{ij}} + \beta \frac{w_s^{ij}}{q_a^{ij}} + \gamma \frac{s_b^{ij}}{s_a^{ij}},$$

q_s^{ij} - количество голосований, на которых оба пользователя i и j сделали одинаковый выбор,

q_a^{ij} - суммарное количество голосований, в которых принимали участие оба пользователя i и j,

w_s^{ij} - взвешенное количество голосований, на которых оба пользователя i и j сделали одинаковый выбор,

s_b^{ij} - количество голосований, на которых оба пользователя i и j отказались голосовать,

s_a^{ij} - суммарное количество голосований, на которых отказался голосовать хотя бы один из пользователей i и j.

Далее к полученной матрице схожести пользователей применялась FLAME кластеризация, разбивающая пользователей на группы со схожими вкусами.

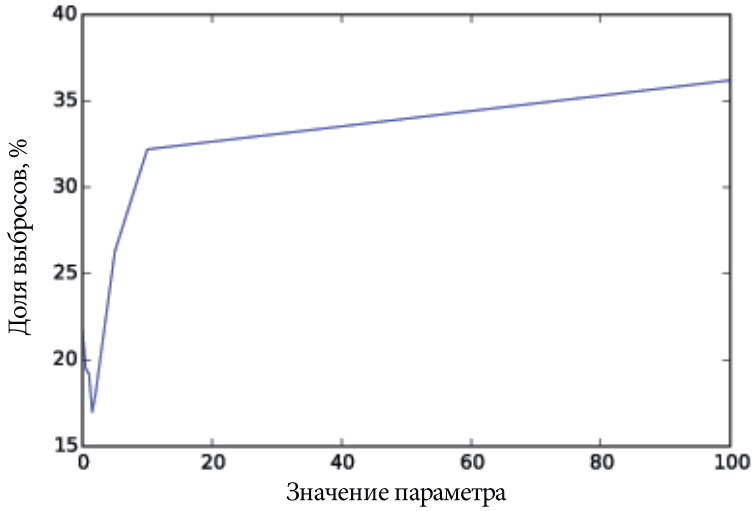
После этого для каждого А/Б-тестирования из тестовой выборки и для каждого кластера пользователей вычислялась доля выбросов, т.е. процент пользователей данного кластера, проголосовавших не так, как большинство пользователей данного кластера. Чем меньше доля выбросов, тем лучше подобраны коэффициенты значимости.

6 Вычислительный эксперимент

Был проведем ряд измерений для различных значений коэффициентов значимости признаков.

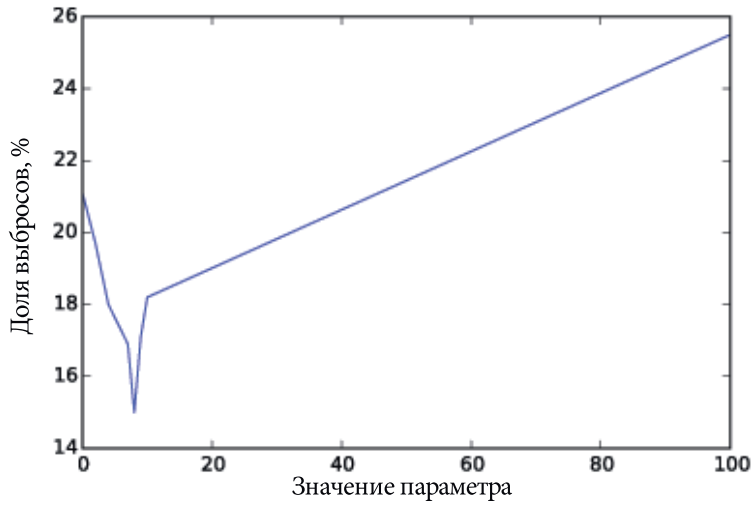
Сначала коэффициенты значимости изучались по отдельности: Доля выбросов для слагаемого α получилась 37%, для слагаемого $\beta - 26\%$, для слагаемого $\gamma - 24\%$.

Далее коэффициенты изучались в комбинации. Фиксировались 2 коэффициента и варьировался 3-й. Сначала изучался коэффициент значимости α при фиксированных $\beta = 10$ и $\gamma = 10$.



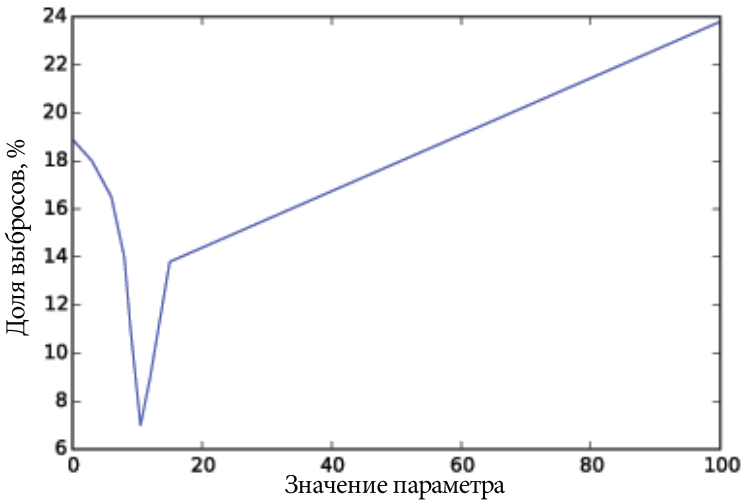
(a) Оптимальный параметр α

Далее изучался коэффициент значимости β при фиксированных $\alpha = 1,5$ и $\gamma = 10$



(a) Оптимальный параметр β

Далее изучался коэффициент значимости γ при фиксированных $\alpha = 1,5$ и $\beta = 8$



(a) Оптимальный параметр γ

7 Основные выводы

Из полученных результатов видно, что

- Ответ пользователей на опросы обусловлен их вкусами ($\alpha = 1,5$)
- Чем уникальнее ответ на опрос, тем больше это говорит о вкусе пользователя ($\beta = 8$)
- Отказ от голосования сильно свидетельствует о вкусе пользователя ($\gamma = 10,5$)

8 Заключение

Вкусы пользователей могут быть выявлены без использования их соц-демографической информации. В результате проделанной работы скорость получения ответа пользователем увеличилась в более, чем 8 раз, что помогло более, чем 400.000 людей, являющимися пользователями сервиса. В будущем планируется совершенствовать функцию похожести и оптимизировать параметры, которые не были затронуты в данной работе.

Литература

1. *Kwan Hui Lim and Amitava Datta* Following the Follower: Detecting Communities with Common Interests on Twitter. – School of Computer Science and Software Engineering. The University of Western Australia

2. *Yang ZHANG, Yao WU, Qing YANG* Community Discovery in Twitter Based on User Interests. – Journal of Computational Information Systems, 2012. Vol. 8, No. 3. P. 991–1000

3. *William W. Cohen* Collaborative Filtering: A Tutorial. – Center for Automated Learning and Discovery. Carnegie Mellon University

4. *Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan* Collaborative Filtering Recommender Systems. – Computer Interaction, 2010. Vol. 4, No. 2. P. 81–173

5. *Jun Wang, Arjen P. de Vries, Marcel J.T. Reinders* Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion.

6. *Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth* User Interests Identification on Twitter Using a Hierarchical Knowledge Base

7. *Ilham Esslimani, Armelle Brun, Anne Boyer* A collaborative filtering approach combining clustering abd navigational based correlations. – KIWI Team, Universite Nancy, LORIA

8. *Subhagata Chattopadhyay, Dilip Kumar Pratihar, Sanjib Chandra De Sarkar* A Comparative Study of Fuzzy C-Means Algorithm and Entropy-Based Fuzzy Clustering Algorithms. – Computing and Informatics, 2011. Vol. 30. P. 701–720

9. *Bahman Bahmani, Andrea Vattani, Sergei Vassilvitskii* Scalable K-Means++

10. *Yang ZHANG* Community Discovery on User Interests

11. *Manh Cuong Pham, Yiwei Cao, Ralf Klamma, Matthias Jarke* A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis. – Journal of Universal Computer Science, 2011. Vol. 17, no. 4. P. 583-604