Метод парных сравнений в задаче нахождения пользовательских предпочтений

А.А. Бородинов[1], В.В. Мясников[1,2]

[1]Самарский национальный исследовательский университет имени академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. В работе рассматривается задача реконструкции функций, заданных неявно результатами парных сравнений значений функций. В предложенном подходе осуществляется переход в пространство большей размерности с последующей классификацией прецедентов-сравнений. В работе в качестве алгоритмов классификации рассмотрены линейная регрессия и случайный лес. Проведены экспериментальные исследования по оценке эффективности предложенного подхода. Показано, что предложенный метод позволяет эффективно решать задачи оценки функции предпочтения пользователей.

1. Введение

Одним из методов, используемых в рекомендательных системах, является метод парных сравнений. Анализируя парные сравнения, мы пытаемся определить некоторую закономерность в выборе предпочтительного варианта. В отличие от классических методов машинного обучения, метод парных сравнений использует информацию о сравнении пар объектов, а не данные о конкретном объекте [1-4]. Задача предоставления рекомендаций для конкретного пользователя системы является задачей выявления предпочтений.

Согласно различным типам наблюдаемой информации выделяют три основных типа задач [3]:

- ранжирование меток – поиск предпочтительного упорядочивания среди меток для любого примера. Традиционная задача классификации может быть обобщена в рамках задачи ранжирования меток, когда результатом классификации примера является метка высшего ранга;
- ранжирование примеров – ранжирование набора примеров для фиксированного порядка меток;
- ранжирование объектов – аналогично ранжированию примеров, однако метки не связаны с примерами.

В данной работе мы рассматриваем задачу ранжирования объектов, где объектами могут быть маршруты, предложенные рекомендательной системой, а предпочтениями являются выбранные пользователем маршруты. Во втором разделе работы мы кратко опишем существующие подходы к построению рекомендательных систем. В третьем разделе приведем основные обозначения и постановку задачи. В четвертом разделе описан метод парных

сравнений. В пятом разделе указаны результаты экспериментальных исследований. В завершении работы представлены выводы и возможные направления дальнейших исследований.

2. Существующие подходы

Существует большое исследовательское сообщество, ориентированное на рекомендательные системы с широким спектром задач. Исторически, большая часть исследований строилась на подходах коллаборативной фильтрации, с упором на прогнозирование рейтингов для стриминговых сервисов, таких как Netflix [5]. Для поисковых систем чаще других применялись попарные методы пользовательских предпочтений [6]. Еще одной крупной сферой применения являются рекомендательные системы, опирающиеся на информации о переходах между сайтами и продуктами в онлайн магазинах [7,8]. Одним из новых подходов стало использование нейронных сетей для повышения точности рекомендаций [9]. Одним из молодых и малоразвитых направлений являются транспортные рекомендательные системы [10]. В своей работы мы рассматриваем метод парных сравнений, который не использовался прежде для построения транспортных рекомендательных систем.

3. Постановка задачи

Рассмотрим множество $\Omega \equiv \{\omega_j\}_{j \in J}$ объектов на котором задано отношение порядка «$\leq$ » и/или строго частичного порядка «$\prec$ ». Эквивалентной будем считать запись $\omega_i \succ \omega_j$ и $\omega_j \prec \omega_i$, а в случае $\omega_i \leq \omega_j \wedge \omega_j \leq \omega_i$ объекты считаем неотличимыми и пишем $\omega_i \sim \omega_j$. Абсолютное предпочтение характеризует функция полезности $u : \Omega \to R$, а относительное предпочтение описывает функция предпочтения $p : \Omega \times \Omega \to R$. Для функции полезности $u(\omega_i) < u(\omega_j)$ можем записать как $\omega_i \prec \omega_j$, $u(\omega_i) \leq u(\omega_j) \Leftrightarrow \omega_i \leq \omega_j$ и $u(\omega_i) = u(\omega_j) <=> \omega_i \sim \omega_j$. А для функции предпочтения $p(\omega_i, \omega_j) > 0$ можем записать как $\omega_i \succ \omega_j$ и $p(\omega_i, \omega_j) = 0 <=> \omega_i \sim \omega_j$. На функцию предпочтений накладываются ограничения, следующие из свойств соответствующих отношений порядка такие как асимметричность по аргументу, транзитивность и т.д.

Функция предпочтения может быть выведена через функцию полезности $p(\omega_j, \omega_i) = u(\omega_j) - u(\omega_i)$ и наоборот $u(\omega_j) = p(\omega_j, \omega^*)$ $\left(u(\omega^*) = p(\omega^*, \omega^*) = 0\right)$.

Объекты описаны вектором признаков $\mathbf{x} \equiv \mathbf{x}(\omega) \in X$ N-мерного пространства. Функцию полезности и предпочтений будем записывать в виде $p(\mathbf{x}, \mathbf{x}_j), u(\mathbf{x})$. Примем $\mathbf{x}_j \equiv \mathbf{x}(\omega_j)$ и $p_{ij} \equiv p(\omega_i, \omega_j)$, $u_j \equiv u(\omega_j)$ для сокращения записи. Информация о парных сравнениях может быть представлена в виде значений функции предпочтения $p(\omega_j, \omega_i)$ либо в виде знакового представления:

$$z_{ij} \equiv z(\omega_j, \omega_i) = \begin{cases} 1, & p(\omega_j, \omega_i) > 0, \\ 0, & p(\omega_j, \omega_i) = 0, \\ -1, & p(\omega_j, \omega_i) < 0. \end{cases}.$$

Примером информации в виде парных сравнений в транспортной рекомендательной системе может служить выбор конкретного маршрута движения среди списка предложенных системой маршрутов.

В качестве критерия качества реконструкции функции предпочтений и полезности будем считать число неверно реконструированных отношений, а именно расстояние Кендалла для парных сравнений:

$$d = \left| \left\{ (i,j) : z(\omega_i, \omega_j) \neq z\left(\mathbf{x}(\omega_i), \mathbf{x}(\omega_j)\right), \quad (i,j) \in I \right\} \right|,$$

значение которого в нормированного виде является оценкой соответствующей вероятности ошибок в отношениях $\tilde{d} = d \cdot |I|^{-1}$.

4. Методология

4.1 Метод парных сравнений

Запишем частоту предпочтения *i*-го объекта над *j*-ым в виде матрицы (c_{ij}) [11]. Для анализа данных в матрице будем использовать модель Терстоуна [12], в которой предполагается, что полезность объекта определяется случайной величиной с нормальным законом распределения. Таким образом для объектов ω_0, ω_1 функцию плотности вероятности запишем как

$f_u\left(u \middle| \omega_j\right) \sim \mathrm{N}\left(\mu_j, \sigma_j^2\right)$, что при $u(\omega_1) - u(\omega_0) \sim \mathrm{N}\left(\mu_1 - \mu_0, \sigma_{10}^2\right)$, $\sigma_{10}^2 \equiv \sigma_1^2 + \sigma_0^2 - 2\rho_{10}\sigma_1\sigma_0$ и функции Лапласа $\Phi(\cdots)$ можно представить следующим образом:

$$P(\omega_1 \succ \omega_0) = P\left(u(\omega_1) - u(\omega_0) > 0\right) = \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_{10}}\right).$$

4.2 Метод реконструкции функций

При реконструкции функции полезности и функции предпочтений следует учитывать следующие особенности:

- при малом объеме информации или при его отсутствии, как в случае холодного старта системы, реконструкция функций практически невозможна;
- следует использовать нелинейные модели, путем перехода в новое признаковое пространство Y большей размерности для разделимости классов почти наверное;
- регрессионная задача реконструкции функции полезности может быть сведена к задаче классификации за счет реконструкции знакового представления.

Метод реконструкции функций по их знаковому представлению опишем следующими этапами:

– нормализация диапазона признаков в [0,1];

– перевод существующего описания – вектора $\mathbf{x}$ в новое пространство признаков Y большей размерности K=dim(Y) ≥ N;

– построение классификатора в новом пространстве Y;

– оценка качества построенного классификатора на тестовом множестве.

В случае, когда оценка функции предпочтения неудовлетворительная, возможен переход к этапу перевода существующего описания в новое пространство признаков с изменением базиса или размерности.

Описанные этапы можно представить в виде схемы, как показано на рисунке 1.

4.3 Банк базисов

Для осуществления отображения $\varphi : \underset{\mathbf{x}}{X} \to \underset{\mathbf{y}(\mathbf{x})}{Y}$ мы используем пять базисов:

– базис исходного представления:

$$K = \dim(Y) = \dim(X) = N, \quad y_n = x_n, \quad n = \overline{0, N-1};$$

– степенной (полиномиальный) базис:

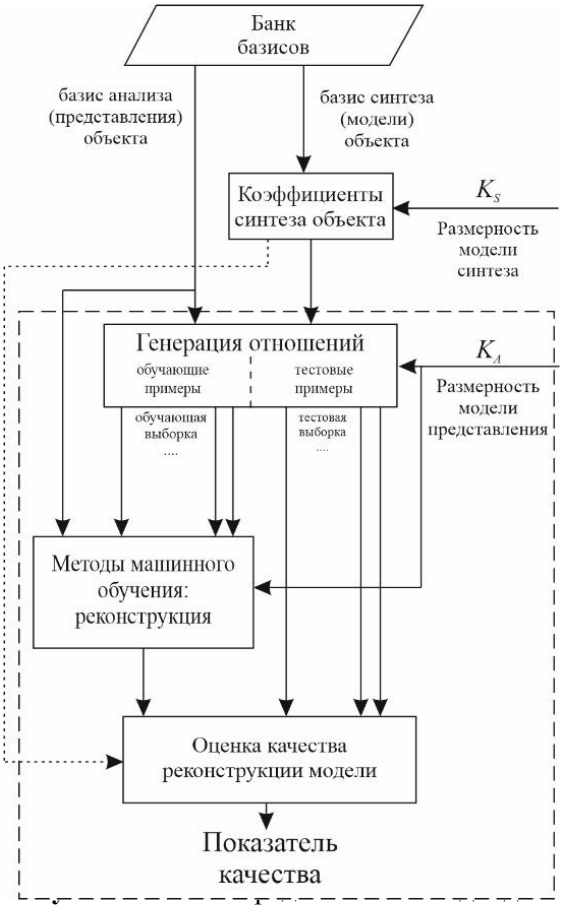
$$K = \sum_{n=0}^{N-1} K_n = [K_n = K_0] = K_0 N = \dim(Y) > \dim(X) = N,$$

$$y_k = \prod_{n=0}^{N-1} x_n^{k_n}, \quad n = \overline{0, N-1}; \quad k = \sum_{n=0}^{N-1} K_0^{\,n} k_n;$$

– Фурье- базис (гармонический):

$$K = \sum_{n=0}^{N-1} K_n = \left[K_n = K_0 \right] = K_0 N = \dim\left(Y\right) > \dim\left(X\right) = N,$$

$$y_k = \prod_{n=0}^{N-1} \cos\left(\pi k_n x_n\right), \quad n = \overline{0, N-1}; \quad k = \sum_{n=0}^{N-1} K_0^n k_n.$$



- Хаар базис:

$$K = \sum_{n=0}^{N-1} K_n = \left[K_n = K_0 \right] = K_0 N = \dim\left(Y\right) > \dim\left(X\right) = N,$$

$$y_k = \prod_{n=0}^{N-1} \varphi(x_n), \quad \varphi(x_n) = \sqrt{2^{k_n}}\,\varphi(2^j x - i),$$

$$i = 0,1,\ldots,2^j - 1, \quad j = 0,1,\ldots,\log_2 N - 1,$$

$$n = \overline{0, N-1}; \quad k = \sum_{n=0}^{N-1} K_0^n k_n.$$

4.4. Методы машинного обучения

В данной работы мы используем логистическую регрессию и случайный лес при тестировании предложенного подхода.

Логистическая регрессия решает задачу бинарной классификации с использованием линейной разделяющей гиперплоскости:

$$d\left(\mathbf{x}\right) = \mathbf{w}^T \mathbf{x} + w_N .$$

Параметры классификатора для конкретной обучающей выборки $\{\mathbf{x}_j, r_j\}_{j\in J}$ определяются из условия:

$$J(\mathbf{w}) = \sum_{j\in J} \ln\left(1 + \exp\left(-r_j \cdot d\left(\mathbf{x}_j\right)\right)\right) \to \min,$$

где $r_j \in \{-1,1\}$ – случайная переменная правильной классификации, определяющая истинный класс соответствующего *j*-го объекта.

Случайный лес (Random Forest) представляет собой реализацию метода голосования нескольких классификаторов-деревьев. В отличие от дерева решений, случайный лес позволяет избежать переобучения. Каждое дерево стоится независимо от остальных на случайном подмножестве обучающего множества. При обучении деревьев для каждого разбиения выбираются компоненты вектора признаков из случайного подмножества признаков.

Пользователи могут делать несвойственный им выбор, особенно при малом различии предложенных альтернатив. Чтобы учесть подобное поведение пользователей мы используем модель Терстоуна для внесения погрешностей в истинные предпочтения. Для случая $\mu_j > \mu_i$:

$$z_{ij} \Leftarrow \begin{cases} z(\omega_j, \omega_i), & rnd < P\left(u(\omega_j) - u(\omega_i) > 0\right), \\ -z(\omega_j, \omega_i), & иначе. \end{cases}$$

Где $rnd \sim R[0,1]$ - случайная величина.

Чтобы избежать влияние неудачного разбиения выборки на обучающую и тестовую, мы проводим обучение и тестирование несколько раз, усредняя результаты расчета ошибок.

5. Экспериментальные исследования

Во время проведения экспериментов были использованы следующие параметры:
– Размерность модели синтеза *Ks* = 15, 35;
– Размерность модели представления *Ka* = 15, 35, 63;
– Число парных сравнений InstNum = 10000, 50000;
– Число разбиений выборки данных на обучающую и тестовую nIter = 100.
Сравнение эффективности различных базисов представлено в таблице 1.

Таблица 1. Результаты сравнения Фурье, степенного и Хаар базисов (LR –логистическая регрессия, RF – Random Forest).

Ks	*Ka*	InstNum	Ошибка											
			S: Фурье A: полином		S: Фурье A: Хаар		S: полином A: Фурье		S: полином A: Хаар		S: Хаар A: Фурье		S: Хаар A: полином	
			LR	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR	RF
15	15	10000	0,2246	0,1129	0,0078	0,0482	0,0092	0,0115	0,0048	0,0101	0,0015	0,0408	0,0068	0,0067
15	15	50000	0,14208	0,06006	0,00172	0,02366	0,01004	0,00952	0,00164	0,0062	0,00222	0,0211	0,00244	0,0048
15	63	10000	0,1986	0,1159	0,0083	0,0509	0,008	0,0139	0,0064	0,0091	0,0067	0,0422	0,0035	0,0051
15	63	50000	0,15958	0,07988	0,00442	0,0319	0,00434	0,00792	0,00144	0,00554	0,00248	0,02138	0,0014	0,00422

Увеличение числа парных сравнений привело к уменьшению значения ошибки, однако значительно увеличило время выполнения программы, особенно для случайного леса. Полученные результаты позволяют утверждать, что предложенный подход продемонстрировал работоспособность и эффективность.

6. Выводы

В работе предложен подход к реконструкции функций, заданных неявно результатами парных сравнений. Подход основан на переходе в пространство признаков большей размерности с последующей классификацией результатов сравнений. Показано, что предложенный метод позволяет эффективно решать задачи оценки функции предпочтения пользователя.

Логистическая регрессия имеет значительное преимущество по скорости и устойчивости. Дальнейшим направлением исследований является применение разработанного подхода на данных о предпочтительных маршрутах движения пользователей на общественном и личном транспорте.

7. Благодарности
Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (уникальный идентификатор проекта RFMEFI57518X0177).

8. Литература
[1] Bradley, R.A. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons / R.A. Bradley, M.E. Terry // Biometrika. – 1952. – Vol. 39(3/4). – P. 324-345.
[2] Фишберн, П. Теория полезности для принятия решений / П. Фишберн – М.: Наука, 1978. – 352 с.
[3] Fürnkranz, J. Preference Learning / J. Fürnkranz, E. Hüllermeier – Berlin Heidelberg: Springer-Verlag, 2011.
[4] Murphy, K.P. Machine Learning: A Probabilistic Perspective. Machine Learning / K.P. Murphy – MIT Press, 2012. – 1098 p.
[5] Koren, Y. Matrix Factorization Techniques for Recommender Systems / Y. Koren, R. Bell, C. Volinsky // Computer. – 2009. – Vol. 42(8). – P. 30-37.
[6] Cao, Z. Learning to rank: From pairwise approach to listwise approach / Z. Cao // ACM International Conference Proceeding Series. – 2007. – Vol. 227. – P. 129-136.
[7] Joachims, T. Optimizing search engines using clickthrough data / T. Joachims // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2002. – P. 133-142.
[8] He, X. Practical lessons from predicting clicks on ads at Facebook / X. He // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.
[9] Covington, P. Deep neural networks for youtube recommendations / P. Covington, J. Adams, E. Sargin // RecSys – Proceedings of the 10th ACM Conference on Recommender Systems, 2016. – P. 191-198.
[10] Campigotto, P. Personalized and Situation-Aware Multimodal Route Recommendations: The FAVOUR Algorithm / P. Campigotto // IEEE Transactions on Intelligent Transportation Systems. – 2017. – Vol. 18(1). – P. 92-102.
[11] Tsukida, K. How to analyze paired comparison data / K. Tsukida, M.R. Gupta. – 2011.
[12] Thurstone, L.L. A law of comparative judgment / L.L. Thurstone // Scaling: A Sourcebook for Behavioral Scientists, 2017. – P. 81-92.

Pairwise comparisons in finding user preferences

A.A. Borodinov[1], V.V. Myasnikov[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. In this paper, we consider the problem of reconstructing functions defined implicitly by the results of pairwise comparisons of the function. In the proposed approach, we map from a low-dimensional space to a high-dimensional. Then we classify the comparisons. In this work, we consider linear regression and random forest as classification algorithms. Experimental studies to evaluate the proposed approach and compare the effectiveness have shown that the proposed method can effectively solve the problem of evaluating the user preference function.