УДК 519.7 DOI: 10.14529/ctcr170402

ОБЪЕКТИВНЫЕ ФАКТОРЫ И ИХ СМЫСЛЫ

Д.В. Гилёв, Вл.Д. Мазуров

Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия

Рассматривается метод построения факторов и новый метод вычисления их смыслов. Предложенный подход к теории факторного анализа тесно связан с комитетными конструкциями, такими как совместные подсистемы неравенств, а также с математической лингвистикой. Рассматривается комитет как коллективное решение. Фактор – латентный источник динамики взаимозависимых признаков объектов и явлений. В одном из популярных методов факторного анализа исследуется структура матриц ковариаций и корреляций. В ниже изложенном подходе используются семантические окрестности признаков. Также в данной статье приводятся элементы математической статистики и распознавания образов. Рассматривается проблема имён и значений факторов как функций имён и значений признаков.

Ключевые слова: факторы, имена, признаки, лингвистика, статистика, алгебра, комитеты.

Введение

Для определения смыслов признаков и факторов мы используем лингвистику как путь к неформальному анализу.

Основоположник факторного анализа – английский исследователь сэр Френсис Гальтон (1822–1911), основатель дифференциальной психологии и психометрии разработал в 1850-е годы исходные идеи факторного анализа с их внедрением в психологическую проблематику индивидуальных различий. Задача состоит в построении математической модели индивидуальных различий. Идея, высказанная Гальтоном, такова: если несколько признаков, измеренных на объектах, имеют согласованную динамику, то следует предположить, что за ними стоят латентные факторы, не имеющие доступных для наблюдений прямых измерений.

В книге [1] рассматриваются математические модели и методы комитетных решений задач распознавания образов, в том числе дискриминантного анализа, таксономии и оценок информативности подсистем признаков. Заметим, что информативные признаки – факторы. Среди комитетных конструкций главные – комитеты большинства и старшинства. Это одна из моделей консилиума экспертов. Основная задача – нахождение решающего правила распознавания образов.

Задача состоит в следующем. Надо найти комитет разделяющих функций для прецедентных множеств A и B. Разделяющая функция f, если она существует, удовлетворяет системе неравенств, которую обозначим через (*):

$f(a) > 0$ для всех a из множества A,

$f(b) < 0$ для всех b из множества B,

f отыскивается в функциональном классе F.

Однако эта система часто бывает несовместной, и тогда вместо одной функции мы строим комитет C функций.

Это конечная последовательность $C = [f_1, \ldots, f_q]$, такая, что каждому неравенству системы (*) удовлетворяют более половины функций из набора C. При этом некоторые из функций набора могут повторяться.

В данной статье изучается связь этих методов с факторным анализом, позволяющим находить глубинные взаимосвязи и смыслы в таблице наблюдений, а также с искусственными нейронными сетями. При этом в отличие от традиционных методов, использующих математическую статистику (которые требуют больших массивов наблюдений и предполагают поиск зави-

симостей признаков от факторов) мы предлагаем алгебраический подход – на основе метода комитетов.

Мы определили направления дальнейшего развития теории и методов исследования операций и распознавания образов. Это методы коррекции решающего правила и нахождение его содержательного смысла через включение лингвистических инструментов.

Один из подходов для такой коррекции связан с построением коллективных обобщённых решений несовместных систем ограничений и опирается на различные логики голосования (демократии), простейшие из которых связаны с принятием решений большинством голосов или с комитетами старшинства.

1. Из истории факторного анализа

Первоисточники комитетной теории можно при желании найти в некоторых американских работах по искусственным нейронным сетям – в алгоритмах Нильса Нильсона, Аблау и Кейлора. Правда, они считали, что нейронные сети относятся к инженерным дисциплинам, и поэтому не ставили перед собой задачи математического строгого обоснования соответствующих алгоритмов.

Имеется целый ряд концепций, исходя из которых строятся решающие правила диагностики и классификации.

Метод коллективных решений нашёл широкое применение в области распознавания образов и классификации объектов и ситуаций, где соответствующие алгоритмы обучения известны под названиями комитетных или ассоциативных (committee machines, associative machines) и усиления – бустинга (boosting). Несмотря на явную близость этих подходов, по ряду причин они долгое время развивались независимо.

М.Ю. Хачай в своей докторской диссертации заметил, что можно осуществить синтез теории комитетов с теорией эмпирического риска.

Сейчас продолжается развитие цикла работ сотрудников ИММ УрО РАН (раньше Мазуров, Тягунов, Казанцев, Кривоногов, Сачков, Белецкий, сейчас Гайнанов, Матвеев, Хачай), ориентированных на выявление глубинных связей между данными подходами, что послужит дальнейшему развитию этих подходов. Так, например, вышла идейно близкая, но совершенно оригинальная и глубокая книга Д.Н. Гайнанова, основанная на комбинаторной геометрии и теории графов.

Развиваются и методы синтеза нейронных сетей на основе метода комитетов.

На примере задачи о минимальном аффинном разделяющем комитете (простейшем кусочно-линейном классификаторе, основанном на голосовании большинством) исследуется теоретико-игровой подход к построению и обоснованию приближённых, в частности полиномиальных, алгоритмов обучения распознаванию и классификации объектов и ситуаций.

Задача построения аффинного разделяющего комитета является дискретным обобщением задачи о разделяющей гиперплоскости в евклидовом пространстве на случай разделяемых множеств, выпуклые оболочки которых пересекаются. Если разделяемые множества конечны, то постановка этой задачи в таком случае естественным образом погружается в конечномерное пространство подходящей размерности.

Один из них использует анализ конечных и потенциально бесконечных систем неравенств – линейных и нелинейных – они могут быть как совместными, так и несовместными. С ним связан и метод комитетов.

Однако метод комитетов принципиально не сводится к разделению двух конечных множеств одной функцией. Он имеет и другие особенности: не требуется выполнения гипотез разделимости, в том числе аксиомы компактности. Но предполагается только выполнение необходимого условия, самого слабого: чтобы обучающие множества разных классов не пересекались. Важно, что при этом минимальном условии всегда существует комитет, состоящий из аффинных функций. Г.Ш. Рубинштейн отметил связь теории комитетов с задачей систем различных представителей набора множеств.

Заметим, что комитет С фактически есть набор факторов.

Другой подход связан с минимизацией эмпирического риска. В.Н. Вапник построил теорию статистических проблем обучения. Он обобщил теорему Гливенко и построил теорию равномер-

ной сходимости частот появления событий к их вероятностям, ввёл меру разнообразия классов функций – vc-коэффициент.

Подход Ю.И. Журавлёва – метод оценок – связан с математическими принципами классификации, этот метод оценок охватывает многие алгоритмы распознавания, в том числе и эвристические. В частности, он строит алгебру алгоритмов, включая эвристические. И в этой алгебре находит оптимальное решающее правило.

О факторном анализе написана необозримая масса книг и статей, и всё-таки сохраняется какая-то особая таинственность этой темы. Дело в том, что требуется содержательный смысл преобразований признаков в факторы. Есть даже обычно совершенно неформальная часть алгоритма. Это назначение смысла фактору, в котором соединены признаки вместе с именами признаков. Так как мы оперируем именами признаков и факторов, то мы используем методы математической лингвистики.

2. Основные понятия и определения

Алгоритм именования приобретает полную формализацию. А именно, пусть X – конечное множество в пространстве Rn. Элементу x из этого пространства ставим в соответствие абстрактную технологию [x; n(x)], где n(x) – имя объекта x, слово в некотором алфавите α. Производим таксономию множества X:

TAXON(X) = {X1, …, Xq}.

Обозначим через V(x) семантическую окрестность элемента x, обращаясь к словарю над алфавитом α, где окрестность – множество синонимов к n(x). Обозначим

$N(X_i) = g(n(x): x \in X_i)$,

это имя фактора, зависящего от X_i.

Ж.-Ф. Лиотар заметил, что иногда мы сталкиваемся с мышлением, которое не является ни философским, ни чисто математическим.

Теперь конкретно об алгоритме вычисления имени фактора по именам признаков, входящих в соответствующий таксон. Мы применяем факторный анализ к таблице наблюдений объект\признак.

Первый этап – построение таксонов столбцов признаков при их пробегании по объектам. В матрице [объекты\признаки]:

строки – величины признаков при просмотре объектов по строке,

столбцы – признаки при их просмотре по объектам.

Метод состоит в следующем. Пусть надо разбить на таксоны конечное множество P в пространстве Rm. И пусть форма таксона задаётся.

Первый этап – построение соответствующей модели.

Второй этап – для каждого таксона (ему соответствует таксон объектов) записать слово из имён признаков. Это слово будет именем фактора.

Третий этап – сжатие большого слова для его преобразования в имя фактора.

Теперь запишем всё это в символьной форме. Матрица наблюдений A представляется двояко – через строки и через столбцы:

$A = [C_1 \ldots C_m]^* = [P_1 \ldots P_n]$.

Здесь C_j – строки, P_i – столбцы, * – знак транспонирования. Обозначим через $a(c_j^*)$ имя объекта, через $a(P_i)$ – имя признака.

Возьмём какой-либо таксон T множества столбцов:

$T = \{P_i: i \in I\}$.

Метод его нахождения заключается в следующем. Пусть P – конечное множество в пространстве Rm. И пусть форма таксона задаётся выражением

$T = \{x: f(x) < 0, x \in P\}$.

Здесь f берётся из допустимого множества F. Отыскивается f из класса функций F.

Таксону T соответствует фактор с именем $[a(P_i): i \in I]$. Это «большое» слово состоит из «малых» слов a(i). Это и есть имя фактора. Можно это слово сжать по мере необходимости. Если w_i – имя i-го признака, а w – искомое имя фактора, то надо найти слово w как функцию $ф(w_i, i = 1, \ldots, n)$. Для этого находим окрестности $V(w_i)$ как множества синонимов. Тогда w при-

надлежит пересечению множеств $V(w_i)$. Численные значения признаков и факторов находятся при решении прямой (ЛП) и двойственной (ЛП*) задач линейного программирования.

Надо заметить, что стандартная интерпретация двойственной задачи ЛП*, когда ЛП – модель экономической задачи – некорректна, потому что получаются нулевые значения некоторых переменных прямой и двойственной задач, а также нулевое значение рентабельности – это обстоятельство не имеет экономического смысла. Но оно естественно при термодинамической интерпретации. Когда состоялась дискуссия математиков с экономистами в 1964 году, то экономисты указывали на это обстоятельство (нулевую рентабельность) как на несогласующееся с экономическим смыслом [2].

3. Смыслы слов и объектов

Лейбниц мечтал о вычислении норм морали. Это было предвосхищением искусственного интеллекта. У меня используется абстрактная модель технологии как элемента в прямом произведении пространства имён на линейное пространство обычных технологий. Эта конструкция даёт возможность вычисления смыслов факторов.

Смысл объекта – это сущность объекта, его место в широком контексте, в реальности. Это и его предназначение, и его значение. Смысл объекта – и его целеполагание, и результат его применения. Его использование зависит и от знания об объекте. Смысл может быть скорее у знакомого объекта. У незнакомого объекта смысл размыт или фантастичен. У древнейших людей возникает придуманный, мифологический смысл. Всё окружающее древнейшего человека имело для него систематическое неотложное значение.

Слова – носители мыслей (текст – длинное слово). Это средства общения, передачи знаний. Всё это касается общей лингвистики. А мы далее будем рассуждать о математической лингвистике.

В эмпирическом смысле первый уровень анализа слова – в его представлении как объекта, второй уровень – нахождение смысла слова.

Факторный анализ требует понимания смысла факторов. То есть мы хотим знать слова как инструменты осмысления результатов математического анализа.

Предполагаем, что заданы: объекты и их имена, признаки и их имена, матрица объекты/признаки. Мы к ней добавим таблицу

(факторы объектов\факторы признаков).

Требуется найти факторы – функции признаков – и их имена, факторы объектов и их имена, по матрице объекты\признаки при этом получаем:

факторы объектов/факторы признаков.

Для формирования факторов применяем таксономию. Для поиска объективно обусловленных имён используем математическую лингвистику. Используется построение данных через абстрактные технологии

(имя технологии/парамеры технологии).

Это делает возможным нахождение объективно обусловленных имён факторов.

4. Общественный договор и факторный анализ

Комитет – это одна из моделей принятия решений коллективом. У медиков это модель консилиума. У экономистов – модель решающей коалиции.

Часто после решения задачи непосредственно дискриминантного анализа встаёт вопрос о выборе в каком-то смысле наилучшего элемента одного из дискриминированного класса. В экономической постановке этот вопрос можно переформулировать так: допустим, мы научились отличать хорошее от плохого состояния рынка. Как мы можем моделировать управляемые параметры, чтобы достичь максимального роста некоторого биржевого индекса, оставляя рынок в хорошем состоянии? Такая задача легко разрешима в случае её сведения к линейному программированию (ЛП).

В случае разделения множеств сложными нелинейными поверхностями, как в методе опорных векторов, такая задача МП сложно решаема.

Теория комитетов построена в трудах Вл.Д. Мазурова и М.Ю. Хачая.

При вычислении имён используются расстояния в словарном пространстве: используются

дисперсные, контекстные, семантические модели, при этом расстояние между словами – количество минимальных операций, превращающих слово в другое слово.

Контекстный вектор для слова w – указание на те слова, вместе с которыми это слово встречается в текстах. Это контекстный вектор первого порядка. Можно построить контекстные векторы второго порядка, третьего и так далее.

5. Элементы математической лингвистики и распознавания образов при агрегировании данных в задачах выбора

Математика – точная наука. Это утверждение неточно. Объектами математики являются идеи. Анализ математических идей происходит на основании принятых в данное время стандартов доказательств. Но, казалось бы, и сугубо содержательные науки – история, филология и даже философия – не могут в наше время не обращаться к математике. Философская логика возрастает на почве математической логики. А сущность математики неуловима, потому что она использует абстрактные модели, но с другой стороны она содержательна, хотя в ней доказательства теорем стараются сделать формальными. Кризисы математики идут ей на пользу.

Не является исключением и лингвистика. Более того, в её разделах есть лингвистические методы распознавания, есть математический структурализм, есть работы пражской русской школа (Якобсон, Трубецкой и другие), есть и философская логика. Среди важнейших приложений математической лингвистики назовём задачу определения авторства произведений, например, идентификацию авторов текстов по распределению букв в текстах. Особенно интересны исследования текстов Шолохова с целью подтверждения его авторства «Тихого Дона». Заметим, что авторство Шолохова подтвердилось в результате комплексной проверки коллектива исследователей.

И есть математическая психология. Не говоря уже о математической экономике, в которой целые разделы науки написаны как совершенно математические, и написаны они математиками. Математика – наука об идеях.

И кроме перечисленных разделов в математике есть теория размытых множеств, есть математическая статистика. Но во всех случаях математические методы не теряют строгости, причём их применения способствуют более глубокому анализу базисных понятий, лежащих в основании как «гуманитарных» наук, так и естественных наук.

В России математическая лингвистика началась с работ А.А. Ляпунова и его школы. Понималась она практически – в русле автоматического перевода с одного языка на другой. В 1954 году Ляпунов стал руководить И.А. Мельчуком – настоящим тонким лингвистом. Но настоящий лингвист не хочет быть формалистом. Математика и глубинная лингвистика трудно совместимы [3]. Но это направление может оказаться главным. Обнаружено, что дело опирается на словари.

Важным разделом математической лингвистики является дистрибутивная семантика. Дистрибутивный анализ возник в 1920-е годы в работах Л. Блумфилда. Развитию этого раздела посвятили свои исследования Ф. де Соссюр и Л. Витгенштейн. В рамках этого направления возникли методы вычисления расстояний между лингвистическими объектами. Дистрибутивный анализ – метод классификации языковых объектов на основе распределения этих объектов в текстах.

Если каждому слову приписать его контекстный вектор, то в целом получится словесное пространство. Контекстные векторы предложены Ч. Осгудом. Семантическое расстояние между понятиями – это расстояние между объектами вербального пространства. Оказалось, что лингвистические объекты, встречающиеся в близких контекстах, имеют близкие значения.

Через контекстные векторы некоторые проблемы лингвистики можно решать как проблемы линейной алгебры.

Для этой темы полезен метод обнаружения супервентности. Супервентность массива A от массива B – это зависимость A от B. В теории линейных неравенств мы рассматриваем условия зависимости одних неравенств от других – это теорема Фаркаша – Минковского.

Мы будем рассматривать метрическое пространство слов. Термин «метрическое пространство» как абстрактное понятие ввёл в 1914 году Ф. Хаусдорф. Количество метрик бесконечно.

Приведём некоторую абстрактную модель. Пусть I – информация, поступающая на вход системы искусственного интеллекта, s – вектор параметров этой системы, t – тезаурус этой системы (это может быть словарь), T – передаваемое сообщение. Тогда информация, воспринимаемая системой, есть некоторая функция ф(I, s, t, T).

А.М. Пешковский назвал лингвистов жрецами грамматической науки. А.Г. Киклевич назвал лингвистику приложением к логике [4].

К данной тематике относятся словарные множества и функции. Важнейшие классы алгоритмов можно описать в терминах теории слов в некотором алфавите.

Пусть α и β означают слова, записанные в алфавите А, не содержащем этих символов. Тогда $\Delta = \Delta(\alpha, \beta)$ обозначает слово, такое, что сначала выписываем слово α, а потом приписываем к нему слово β, и получается композиция двух слов. Операция композиции слов ассоциативна, но не коммутативна. И дальше рассматривается словарное множество над алфавитом А.

В факторном анализе необходимо давать объяснение результатов, поэтому мы используем математическую лингвистику, встраивая её процедуры в алгоритмы обработки данных и знаний. Объекты языка образуют систему словарей, связанных сложными сетями отношений.

Рассмотрим проблему имён и значений факторов как функций имён и значений признаков. Имени признака соотнесём множество его синонимов. Причём слово является синонимом самого себя. И тогда имя фактора – пересечение этих множеств. К этому множеству мы добавляем слова, близкие по расстоянию строк и перестановок. По этой же схеме результаты работы алгоритмов надо истолковывать и именовать. Теоретическая интерпретация – смысл знакового абстрактного выражения объёмов понятий.

Введём дальнейшую формализацию. Алфавит – это конечное множество А, состоящее из не менее чем из двух элементов. Элементы этого множества – буквы или символы. Строка или слово – это последовательность символов из конечного алфавита А. Множество всех конечных последовательностей над алфавитом А обозначим через W(A).

В одном из толкований факторов это подстрока пространства строк

$x = x_1 \ldots x_n$

– это любая подпоследовательность смежных элементов строки. В другом значении фактор – это скрытая причина, сжатое сочетание признаков. Префикс строки – любая её подстрока, начинающаяся с символа x_1. Суффикс – любая подстрока, заканчивающаяся на x_n.

Проблема определения значения и смысла слов – предмет семиотики, охватывающей лингвистический и логико-семантический анализ. Можно здесь применить гипотетическую модель реализации свойства предметности чувственного образа.

Мы стремимся к эмпирической интерпретации решений, выдаваемых компьютером. В том числе к интерпретации факторов. В отличие от установившегося понятия фактора как глубинной причины явления мы ищем таксономию признаков, и тогда имя фактора – это функция имён признаков, входящих в соответствующий таксон.

До сих пор в качестве имени фактора применялась такая громоздкая конструкция как кортеж имён признаков, входящих в фактор – такую конструкцию применяет И.Б. Мучник. Мы применяем более подходящую конструкцию, возникающую при работе со словарём. При этом цель факторного анализа – установление общих закономерностей, определяющих сущность исследуемого явления. Имя фактора – слово как элемент пересечения объёмов имён признаков.

Исходная задача – анализ объектов с помощью системы признаков, двойственная – анализ признаков, описание которых делается по их динамике на объектах.

Наконец, также члены разделяющего комитета могут получить объективные имена, если мы применим подобную технику вычисления смыслов уже не для подсистем признаков, а для подсистем объектов.

Работа поддержана РНФ – проект № 14-11-00109.

Литература

1. Мазуров, Вл.Д. Метод комитетов в задачах оптимизации и классификации / Вл.Д. Мазуров. – М.: Наука, 1990. – 348 с.

2. Мазуров, Вл.Д. Модель экономической динамики в противоречивых условиях / Вл.Д. Мазуров // Вестник Уральского института экономики, управления и права. – 2012. – № 3 (20). – С. 88–90.

3. Математическая лингвистика / И.Г. Пиотровский и др. – М.: Высшая школа, 1977. – 383 с.

4. Киклевич, А.К. Динамическая лингвистика. Между кодом и дискурсом / А.К. Киклевич. – Харьков: Гуманитарный Центр, 2014. – 442 с.

Гилёв Денис Викторович, аспирант Института естественных наук и математики, старший преподаватель кафедры эконометрики и статистики Высшей школы экономики и менеджмента, Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, г. Екатеринбург; deni-gilev@ narod.ru.

Мазуров Владимир Данилович, д-р физ.-мат. наук, профессор-консультант департамента математики и компьютерных наук Института естественных наук и математики, профессор кафедры эконометрики и статистики Высшей школы экономики и менеджмента, Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, г. Екатеринбург; vldmazurov@gmail.com.

Поступила в редакцию 8 мая 2017 г.

DOI: 10.14529/ctcr170402

OBJECTIVE FACTORS AND THEIR MEANINGS

***D.V. Gilev**, deni-gilev@narod.ru,*
***V.D. Mazurov**, vldmazurov@gmail.com*
Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, Russian Federation

In this article, we consider the method of constructing factors and a new method for calculating their meanings. The proposed approach to the theory of factor analysis is closely related to commutative constructions, such as joint subsystems of inequalities, and also with mathematical linguistics. The committee is considered as a collective decision. The factor is a latent source of the dynamics of interdependent characteristics of objects and phenomena. In one of the popular methods of factor analysis, the structure of covariance and correlation matrices is studied. In the approach described below, the semantic neighborhoods of the features are used. Also in this article are the elements of mathematical statistics and pattern recognition. The problem of names and values of factors as functions of names and values of attributes is considered.

Keywords: factors, names, signs, linguistics, statistics, algebra, committees.

References

1. Mazurov Vl.D. *Metod komitetov v zadachakh optimizatsii i klassifikatsii* [Committees Method in Problems of Optimization and Classification]. Moscow, Science Publ., 1990. 348 p.

2. Mazurov Vl.D. [Dynamic Economic Model in Contradictory Conditions]. *Bulletin of the Ural Institute of Economy, Control and Right*, 2012, no. 3 (20), pp. 88–90. (in Russ.)

3. I.G. Piotrovskiy et al. *Matematicheskaya lingvistika* [Mathematical Linguistics]. Moscow, Higher School Publ., 1977. 383 p.

4. Kiklevich A.K. *Dinamicheskaya lingvistika. Mezhdu kodom i diskursom* [Dynamic Linguistics. Between Code and Discourse]. Khar'kov, Humanitarian Center Publ., 2014. 442 p.

Received 8 May 2017

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Гилёв, Д.В. Объективные факторы и их смыслы / Д.В. Гилёв, Вл.Д. Мазуров // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2017. – Т. 17, № 4. – С. 13–19. DOI: 10.14529/ctcr170402

FOR CITATION

Gilev D.V., Mazurov V.D. Objective Factors and Their Meanings. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2017, vol. 17, no. 4, pp. 13–19. (in Russ.) DOI: 10.14529/ctcr170402