УДК 658. 1 – 50

Мазуров Владимир Данилович,
доктор физико-математических наук, профессор,
кафедра эконометрики и статистики,
Высшая школа экономики и менеджмента,
ФГАОУ ВО «УрФУ имени первого Президента России Б.Н.Ельцина»
e-mail: vldmazurov@gmail.com
г. Екатеринбург, Россия

ФАКТОРНЫЙ АНАЛИЗ И СМЫСЛ ФАКТОРОВ КАК ФУНКЦИЯ СМЫСЛОВ ПРИЗНАКОВ

Аннотация:
Рассматривается метод факторного анализа, включающий в себя именование факторов. Если x – вектор значений признаков, а f - вектор значений овфакторы, то мы ищем зависимости f(x), а не x(f).

Ключевые слова:
Факторы, имена, признаки, лингвистика, статистика, алгебра.

Работа поддержана грантом РНФ № 14 – 11 -00109.

Наш подход к теории факторного анализа тесно связан с теорией комитетных решений, поэтому я счёл необходимым некоторое внимание уделить истокам этой теории.

Фактор – латентный источник динамики взаимозависимых признаков объектов и явлений.

Основоположник факторного анализа – английский исследователь сэр Френсис Гальтон (1822 – 1911), географ, антрополог, психолог, основатель дифференциальной психологии и психометрии, статистик. Он разработал в 1850 – е годы исходные идеи факторного анализа с их внедрением в психологическую проблематику индивидуальных различий. Она – 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 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. Среди комитетных конструкций главная – комитет большинства. 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 (*):

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

$$C = [f1, \dots fq] ,$$

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

Наши работы поддержаны академиком РАН В. И. Бердышевым и академиком РАН Ю. И. Журавлёвым, который руководит в Российской Федерации в целом направлением по алгебраическим моделям, методам распознавания, их математическим и практическим обоснованиям.

Однако надо начать с инициативы выдающихся математиков С. Б. Стечкина и И. И. Еремина, поставивших в 1965 году передо мною задачу доказательства необходимого и достаточного условия существования комитета системы линейных неравенств.

Фундаментальные результаты академика РАН И. И. Ерёмина в области теории и методов решений и оптимальной коррекции несовместных систем уравнений и неравенств, а также и противоречивых задач эффективного (в частности оптимального) выбора определили направления дальнейшего развития теории и методов исследования операций и распознавания образов.

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.

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 – это инженерная дисциплина, 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 (Мазуров, Тягунов, Казанцев, Кривоногов, Сачков, Белецкий, 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 разделяемыз 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). В. Н. Вапник построил теорию статистических проблем обучения. Он обобщил теорему Гливенко и построил теорию равномерной сходимости частот появления событий к их вероятностям, ввёл меру разнообразия классов функций.

Подход Ю. И. Журавлёва – метод оценок – связан с мате –матическими 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. 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.

При этом алгоритм именования приобретает полную формализацию.

Сюда подходит замечание Ж. - Ф. Лиотара: в Sensus communis мы сталкиваемся с мышлением, которое не является философским или 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 объект\признак: строки – величины признаков при пробегании объектов по строке, столбцы – признаки при их пробегании по объектам.

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

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 А представляется двояко – через строки и через столбцы:

$$A = [C1 \ldots Cm]^* = [P1 \ldots Pn].$$

Здесь С*j – строки, Pi – столбцы, * - знак транспонирования. Обозначим через a(c*j) имя объекта, через a(Pi) – имя признака.

Возьмём какой – либо таксон T множества столбцов:

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

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

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

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

Таксону T соответствует фактор с именем [a(Pi): i € I]. Это «большое» слово состоит из «малых» слов a(i). Это и есть имя фактора. Можно это слово сжать - по мере необходимости. Если wi - имя i - го признака, а w - искомое имя фактора, то надо найти слово w как функцию ф(wi, i = 1,…,n). Для этого находим окрестности v(wi) как множество синонимов. Тогда w принадлежит пересечению множеств v(wi). Численные значения признаков и факторов находятся при решении прямой и двойственной 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 подходит к термодинамической интерпретации. Когда состоялась дискуссия математиков с экономистами в 1964 году, то экономисты указывали на это обстоятельство, но тогда математики были в фаворе.

Заметим, что факторный анализ начался с анализа психологических исследований. По – видимому, самая ранняя работа в этом направлении принадлежит К. Пирсону – в 1901 году он публиковал статью «On lines and planes of closest fit to system of points in space», в ней обсуждалась идея главных осей. Далее Ч. Спирмен в 1904 году опубликовал статью «Общий интеллект, объективно определённый и измеренный» в «Американском психологическом журнале». Психодиагностические тесты Г. Роршах ввёл в 1921 году.

Факторный анализ в социологии начался в 1940 году.

Список использованных источников

1. Вл. Д. Мазуров. Метод комитетов в задачах оптимизации и классификации – М. – «Наука» - 1990.
2. Вл. Д. Мазуров. The factor analysis and the search for objective meaning of factors as the function of names features. - Bulletin Of the South Ural University - 2016 - Vol. 16 - no. 3.
3. И. И. Ерёмин, Вл. Д. Мазуров, Н. Н.Астафьев. Несобственные задачи линейного и выпуклого программирования – М. –«Наука» - 1983.

4. Вл. Д. Мазуров. Обобщение комбинаторного метода факторного анализа + таксономия – Вестник ЮУрГУ\ серия «Компьютерные технологии, управление, радиоэлектроника – 2015, том 15, № 2.
5. Ф. – Ж. Лиотар. Содержание постмодерна. – изд. Аксиома – Спб – 2001.
6. А. А. Любищев. Таксономия. – в кн.: А.А. Любищев, Линии Демокрита и Платона в истории культуры – изд -– во Алетейя - Спб – 2001.
7. Н. С. Енюков (ред.) Факторы, дискриминация, кластерный анализ. – М. – Наука - 1982.
8. А. А. Марков. Введение в теорию кодирования – М. – «Наука» - 1982.
9. В. М. Жуковская. И. Б. Мучник. Факторный анализ в социально – экономических исследованиях. – М. – Статистика – 1976.
10. А. А. Марков. Введение в теорию кодирования. – М. – Наука – 1982.
11. Н. Хомский. Синтаксические структуры. – в сб. «Новое в лингвистике» - вып. 2 – М. – 1962.
12. Б. М. Каганович, С. П. Филиппов. Равновесная термодинамика и математическое программирование. - Новосибирск - Наука - 1995.
13. Л. Лопатников (ред.) Экономисты и математики за круглым столом. - М. - Экономика - 1965.
14. М. Ю. Хачай – Докторская диссертация – ВЦ РАН – 2004.
15. Д. Н. Гайнанов. Комбинаторная геометрия и графы в анализе несовместных систем и распознавания образов. – М. – Наука – 2014.
16. Джон Пассмор. Структура и синтаксис. – в книге:
17. Дж. Пассмор. Современные философы – Идея – Пресс – М. – 1982.
18. Дж.- О. Ким, Ч. У. Миллер. Факторный анализ: Статистические методы и практические вопросы. – Eleventh Printing – 1986.
19. M. Dummeth, M. Frege. In: M. Platts, Ways of meaning – 1997.
20. Вл. Д. Мазуров (ред.) Метод комитетов в распознавании образов – Свердловск – УНЦ АН СССР -1974.
21. Ф. П. Чернавин. Моделирование и расчёт кредитного риска заёмщика с применением комитетных конструкций – кандидатская диссертация – УрФУ – 2016.
22. Ю. И. Журавлёв и др. О математических принципах классификации предметов и явлений. – Дискретный анализ. – сб. трудов ИМ СО АН СССР. – Новосибирск – 1966.
23. В. Н. Вапник. . Теория распознавания образов. – М. – «Наука» - 1974.
24. Ю. И. Журавлёв (ред.). – Распознавание -
25. Классификация – Прогноз – М. – «Наука» - 1989.-
26. Вл. Д. Мазуров, В. С. Казанцев, Н. Г. Белецкий, А. И. Кривоногов, А. И. Смирнов. Вопросы обоснования и применения комитетных конструкций распознавания. – Там же.

27. А. А. Белых. История росийских экономико – математических исследований – М. – Urss – 2007.
28. V. I. Glivenko. Sulla determinatione empirica di probabilita. – Giornale della Institute Italiano degli Attuari, 4, 1933.
29. Н. Н. Астафьев. Линейные неравенства и выпуклость. – М. – Наука – 1982.
30. Г.У. Кун, А. У Таккер (ред.) Сб. Линейные
31. неравенства и смежные вопросы – М. – ИЛ -1858.
32. Д. Н. Гайнанов. Разделение пространства выпуклыми конусами. – в Сб. Комбинаторные свойства выпуклых множеств и графов. – Свердловск – 1987.
33. Н. Нильсон. Обучающиеся машины. – М. – «Мир» - 1968.
34. А. О. Матвеев. Комплексы систем представителей в исследовании комбинаторных свойств частично упорядоченных множеств и несовместных систем линейных неравенств. – кандидатская диссертация – Екатеринбург – 1994.
35. P. Hall. On representatives of subsets – Journ. Lond. Math. Soc., 10 (1935).
36. Г.Ш Рубинштейн. Линейное програмирование в СССР. – в сб. Лин. нер – ва и смежные вопросы – М. – Мир – 1959.
37. Б. Д. Парыгин (ред.) Проблемы философии и социологии – Л. – 1968.

Vladimir Mazurov,
doctor of Physical and Mathematical Sciences, professor,
department of Econometrics and Statistics,
Graduate School of Economics and Management,
Ural Federal University named after the first President of Russia Boris Yeltsin
e-mail: vldmazurov@gmail.com
Ekaterinburg, Russia

THE FACTOR ANALYSIS AND SEARCH OF FACTORS OBJECTIVE SENSE LIKE SOME FUNCTION OF SIGNS SENSES

Abstract:
The factor analysis is considered including the factors names.

Keywords:
Factors, names, linguistics, statistics, algebra.