Б.Г. Миркин

К истории развития анализа данных в России

Доклад на заседании кафедры 5 декабря 2012 г. (по случаю 70-го дня рождения)

План

a. Вводные замечания: роль биографии
b. Список российских достижений и пояснения
c. Мои собственные разработки
d. Выводы

Не могу аккуратно отделить историю развития от своей биографии, т.е. куски науки и биографии будут перемежаться.

Я родился и в значительной степени сформировался в сталинскую эпоху. В частности, большое влияние на меня оказала послевоенная кампания по установлению приоритета российской науки и инженерии в развитии научно-технического прогресса. Там были и панегирики русским инженерам, в какой-то мере участвовавшим в создании той или иной технологии: Попов (радио), Лодыгин (электрическая лампочка), Можайский (самолёт) и пр. Бывала и прямая ложь, если не в формулировках, то в предполагаемой интерпретации. Особенно популяризировались отец и сын Черепановы, якобы создавшие поезд и железную дорогу где-то на Урале впервые в мире. Оказалось, что барин отправил их в Англию, чтобы освоить это новое дело, и они по возвращении, действительно построили дорогу, но не впервые в мире, а по изученным английским образцам. Тем не менее я, занимаясь разработкой методов анализа данных, видел, что и я, и мои коллеги делаем что-то абсолютно новое и не так уж и зависим от международной науки. Поэтому я был очень озадачен, когда позже где-то в 1993 году читал купленную мною на развале книгу «Все о мире» - гигантский справочник, напечатанный очень мелким шрифтом. В частности, там был список основных изобретений человечества и их авторов – из этой тысячи только одна фамилия была русская: Менделеев, периодическая таблица. Сначала я очень огорчился таким явным пренебрежением к творчеству Российского народа и посчитал список ещё одним проявлением холодной войны. Теперь я смирился. При всем нашем «хвалёном» творческом подходе у нас никогда не хватает обстоятельств для создания технологии, а только такие изобретения идут в копилку человечества. Что далеко ходить? При нашем общепризнанном программистском потенциале, Россия даёт в мировую копилку не более 1%.

Сейчас я так понимаю развитие науки и техники после войны – США несомненный лидер, все остальные вносят посильный вклад, но вклад является действительно вкладом в мировую копилку, только если он используется в США.

Первоначальное развитие методологии и методов анализа данных в России, как и в Советском Союзе в целом, в целом следовало международному. Несмотря на эффективный железный занавес, препятствовавший свободному обмену, в СССР переводились отдельные монографии, иностранные

журналы достигали библиотеки, хотя и немногие, и с отставанием в год-два, да и отдельные советские научные сотрудники участвовали в некоторых конференциях. Международные подходы к факторному анализу, распознаванию образов и кластер-анализу доминировали. Как самостоятельная дисциплина анализ данных не выделялся и, более того, тогда, да и до сих пор, рассматривался как некая деятельность, к науке отношения не имеющая. В то же время разрабатывались и оригинальные подходы. Упомяну те из них, в которых приходилось разбираться, как говорится, «с карандашиком».

Наиболее популярной оказалась двойная разработка Института проблем управления, действительно вошедшая в мировую копилку: Минимизация эмпирического риска (Вапник-Червоненкис, 1974) и метод обобщенного портрета (SVM, Вапник-Лернер, 1964 – к сожалению, А.Я. Лернер выбыл из круга науки; где-то в начале 70 он подал на выезд и СССР и просидел в «отказе», а значит, и вне общественной жизни, лет 17), особенно обрамленные в так называемые кернел-функции (потенциальные функции – Мучник, Браверман, Розоноэр, Айзерман 1964-1970). Это, безусловно, явление мирового порядка; а кернел-функции - это такое, интенсивно изучаемое понятие, которое еще только предстоит освоить.

Другие оригинальные разработки периода 65-85 годов не получили столь широкого признания. Среди них упомяну следующие (следуя статье Mirkin, Muchnik 2008):

- М.М. Бонгард предложил логические методы анализа данных, включая изобретённые им примеры анализа простейших геометрических образов (Бонгард 1967/1970). Надо отметить, что эти примеры получили широкую известность в международной когнитивной психологии как проблемы Бонгарда (см., например, Linhares (2000));

- Ю.И. Журавлев разработал систему количественных характеристик минимальных подмножеств в бинарных таблицах, так называемых «тупиковых тестов» (Журавлев и Юнусов 1971) и соответствующий алгебраический формализм (Журавлев 1978). Подход хорошо работает на практических задачах -= жалко, что это направление осталось как-то в стороне от магистрали международного развития.

- Н.Г. Загоруйко и Е.Е. Витяев предложили формально-логический аппарат, включающий вывод закономерностей из наблюденных эмпирических данных (Загоруйко 1979). Эта работа осталась, факически, незамеченной, а в ней содержится очень интересная философия и работающий подход, совместимый с мат.статистикой;

- Г.С. Лбов предложил использовать логические решающие правила для данных, измеренных в разнотипных шкалах, во всех направлениях анализа данных (Лбов 1979), включая очень оригинальный метод кластер-анализа, который ещё ждет своей проработки;

- Б.Г. Миркин предложил использовать геометрическое пространство разбиений в качестве инструмента анализа категоризованных данных (Миркин 1969), позднее распространенное на случай наличия разнотипных шкал (Миркин 1985);

- Б.Г. Миркин и И.Б. Мучник разработали, в творческом соперничестве, подход к аппроксимации больших социальных сетей малыми графами на разбиениях (Миркин (1974, 1976, 1981), Мучник (1974), Мучник и Ослон (1980), Браверман и Мучник(1983)). Эта тематика соприкасалась с так называемыми блок-моделми

международной социальной психологии. В настоящее время, по моему, имеется потребность в визуализации сетевых структур, в которой эти подходы могли бы быть применены;

- И.Э. Муллат и И.Б. Мучник разработали оригинальный класс просто оптимизируемых ординальных «квази-выпуклых» функций множеств для отыскания представительных вложенных кластеров в организационных и других системах (монотонные системы Муллат (1976), Кузнецов и Мучник (1983));

- И.Б. Мучник и Э.М. Браверман (1971) предложили метод «лингвистического анализа», предшественник современных методов, становящихся все более популярными, “subspace clustering”, для агрегирования больших матриц данных;

- А.И. Орлов разработал основы вероятностного анализа неколичественных данных (Орлов 1979);

- Л.А. Растригин и Р.Х. Эренштейн предложили использовать коллективы решающих правил для более точного распознавания (Растригин и Эренштейн 1978). Сейчас эта тематика пользуется огромной популярностью, так как действительно, в принципе, позволяет строить все более точные классификаторы;

- В.С. Файн предложил использовать методы непрерывной теории групп для описания геометрических преобразований элементов изображений (Файн 1970). Эта разработка значительно опережала свое время – только сейчас появляются средства автоматизации анализа поворота головы на изображениях, которые так удачно были охвачены в работе Файна. Можно только пожалеть, что эта деятельность как-то сошла на нет, возможно, из-за неудач Файна с защитой его докторской.

- В.К. Финн предложил метод ДСМ (по имени английского философа и экономиста Джона Стьюарта Милля), основанный на отображении сходства между двумя комплексными объектами с использованием всех подмножеств (под-структур) в пересечении этих объектов (Финн 1983);

- С.В. Чесноков разработал так называемый детерминационный анализ на основе условных вероятностей (Чесноков 1982), предваривший анализ ассоциаций в разработке данных (data mining);

Каждое из этих направлений – а также многие другие, здесь не упомянутые, разрабатывалось коллективами единомышленников и породило много статей, в некоторых случаях несколько десятков, публиковавшихся в основном в малотиражных выпусках трудов. Помимо невысокого качества бумаги и печати, эти выпуски обычно посвящались широкому кругу вопросов науки и техники, и конечно, бывали доступны только узкому кругу личных знакомых автора. Вместе с тем, работы по анализу данных публиковались некоторыми журналами, из которых следует отметить прежде всего издание Института проблем управления “Автоматика и телемеханика” – в разделе «Моделирование интеллекта и поведения». Регулярные семинары, руководимые С.А. Айвазяном (ЦЭМИ Москва), М.А. Айзерманом (ИПУ Москва) и Н.Г. Загоруйко (ИМ Новосибирск) сыграли существенную роль в развитии некоторых общих задач и подходов. Конференции были случайны и редки. Например, Э.М. Браверман и И.Б. Мучник организовали Семинар по машинному обучению в Калинине (теперь Тверь) 1970 г., а Б. Г. Миркин и Ф.М. Бородкин организовали семинары по анализу данных в социологии (Новосибирск, 1970, Челябинск, 1977, Улан-Удэ, 1979). Фундаментальную организующую роль, однако, сыграли две серии конференций. С.А. Айвазян (ЦЭМИ Москва) организовал конференции по «прикладной статистике», проводившиеся каждые два года попеременно в Эстонии (Тартуский госуниверситет, организатор А.-М.

Тийтс) и Армении (Цахкадзор, организатор В.С. Мхитарян), начиная с 1977. Н.Г. Загоруйко организовал серию совещаний по «Машинному обнаружению закономерностей», каждые три года, начиная с 1976 г.

Теперь я перейду к списку некоторых своих результатов, многие получены совместно с моими сотрудниками в ИЭиОПП СО АН СССР (1967-1982), ЦЭМИ (1983-1991), ДИМАКСЕ (1993-1998), Биркбеке (2000-2010). Прежде всего, два результата вводящие конструкции, на которых с недавних пор я заметил в Гугле своё имя, иногда даже и без ссылок.

1966 – понятие предбазиса для синтеза автоматов через уравнения в алгебре событий (Mirkin's prebase)

1970 – расстояние в пространстве разбиений с аксиоматическим обоснованием (Mirkin distance)

Затем:

- 1973 – обобщение теоремы Эрроу из социального выбора на произвольные бинарные отношения, с последующим переносом на так называемые федерационные правила агрегирования (1978). Последнее было сделано в результате критического анализа работы Б. Монжарде (Франция 1976), которую автор прислал мне сразу по публикации. Монжарде занимался только так называемыми турнирами, отношениями без «безразличий», и для них пришел к теореме, аналогичной теореме минимакса в теории игр. Я обобщил эти построения на произвольные отношения, потерев при этом двойственность, но зато прибретя то, что я назвал «федерацией».
- 1974 – метод агрегирования структур и использование для совершенствования оргструктуры предприятий промышленности
- 1976 – метод качественного факторного анализа матриц связи, реализованный впоследствии через:
  - метод последовательного исчерпания аддитивных кластеров (1987)
  - метод отщепления бикластеров (1995)
  - метод аномальных кластеров (2005)
  - спектрально-аддитивный метод нечеткого кластер-анализа (2009)
- 1990 – метод «главных» кластеров для данных в смешанных шкалах
- 1990 – развитие теории кластер-анализа на основе критерия наименьших квадратов, приведшей, в частности, к установлению явной связи между нормировкой признаков, их вкладом в кластеры и мерами ассоциации между качественными признаками
- 1994 – отображение генного древа эволюции в общее древо эволюции через понятие дупликации как движущего механизма эволюции
- 2003 – метод построения сценариев эволюции отдельных генов (через их потери) на филогенетическом древе и построение LUCA 572 гена.
- 2003 – метод пропорциональных нечетких кластеров
- 2006 – метод распознавания спам-сообщений с помощью аннотированного суффиксного дерева
- 2009 – метод АПП Автоматического Представления Деятельности путем подъема множества запроса по онтологии предметной области
- 2012 – Метод ФАДДИС для выявления нечетких кластеров по матрице связи (спектрально-аддитивный; более эффективный, чем другие)
- 2012 – Интеллигентная версия метода к-средних с автоматическим определением весов признаков для каждого кластера (с использованием метрики Минковского)

Какие же уроки можно извлечь из моего почти 50 летнего опыта работы, борьбы (4 докторские с 1974 по 1990) и странствий (1991-1993 Франция, 1993-1998 США, 1996-1999 Германия, 2000-2011 Англия). Перечислю несколько обобщений, к которым я пришел за это время.

1. Чем заниматься? Здесь разные точки зрения, например, (а) Найти проблему, которой можно заниматься всю жизнь. (Винер) и (б) Каждые 7-10 лет меняйте тему (Айзерман).

Я думаю, здесь есть два аспекта: (1) сама проблема, (2) Ваша роль в работе по ее решению. И то, и другое – в зависимости от Вашей персональной когнитивной системы. Надо понять, в чем твоя сила, и делать именно это, скажем – сама наука или организация науки. Мне приходилось заниматься и тем и другим, и для меня занятия самой наукой решительно лучше, чем ее организацией. Каждый должен правильно проставить приоритеты и затем им следовать. Я лично ставил задачу так - надо выявить структуру в данных, потом плавно перешел на – классификацию и кластер, из-за связи с Журналом и Обществами классификации. Теперь я перешел на задачу интерпретации данных и текстов; причем беру на себя все роли в соответствующих разработках - и швец, и жнец, и на дуде игрец, потому что избегаю быть руководителем – не способен добывать деньги из-за того, что не считаю эту деятельность серьёзной. Следует отметить, что зарубежный опыт научил меня использовать студентов в своей научной работе – ранее вещь для меня невозможная: цикл научного проекта – год, два или больше, тогда как курсовая или выпускная работа – несколько месяцев, причем студент еще и занят повседневными занятиями.

2. Как относиться к неудачам.
    a. Не слишком серьёзно относиться к себе и своим результатам. (Палка о двух концах – если сам себя не ценишь, кто же будет? Пример Маркса очень показателен – при относительно корявой идее, совершенно неадекватной реальным процессам, его суперсерьезное отношение к себе сыграло огромную роль в распространении его «учения».)
    b. Больше думать про будущее, чем про прошлое. Если ты сидишь в прошлом, вздыхая, эх как же я неудачно сказал или сделал тогда то… то это самоедство, оно съедает человека и его движение останавливается.
    c. Не пытаться переделать систему, чтоб доказать правоту. Такие попытки дорого стоят – в процессе борьбы Вы легко можете уйти с научной тропы. Не бояться отступать или идти в обход.
    d. Если чего-то нельзя или не получается, делать что-нибудь смежное.

3. Как публиковаться

    Так, чтобы вписаться в процесс. В этой связи очень неудовлетворительна политика поощрения (А) по числу публикаций (плодит нечитаемые и неинтересные работы и сборники) или по (Б) индексу цитирования (плодит вторичные работы – ни один лауреат премии Тюринга последних лет не имеет значимого индекса). Не могу не упомянуть происходящий в ВШЭ процесс, при котором такие люди как я лишаются публикаций вообще. Создан специальный отдел, который будет правильно оформлять и, как я понял, ещё и проверять публикации. Это значит, что ВШЭ собирается идти по пути (А), что абсолютно противоречит официально прокламируемой стратегии на достижение высокого индекса цитируемости. В этой связи сошлюсь на опыт Соединенного Королевства в борьбе с метрическими показателями. Там каждый включенный в научную часть департамента, представляет всего 4 публикации за 5-6 лет, но зато указывает, чего в них такого сделано, чего не было раньше. Эти работы смотрят и обсуждают специалисты – члены оценочной комиссии и затем уже относят их к категориям серьезности результата.

Заключая свое выступление, дайте мне ещё раз процитировать основную молитву образованного человека.
Господи, дай мне смирение, чтобы принять то, что я не могу изменить, дай силы изменить то, что могу, и мудрость, чтобы отличить одно от другого.

Здесь перепечатка заметки П. Радзиховского с изложением моей лекции, см. http://www.hse.ru/news/recent/68694857.html

Борис Миркин: к решению научных задач

5 декабря ведущему ординарному профессору НИУ ВШЭ Борису Миркину исполнилось 70 лет. По этому случаю в университете состоялась лекция юбиляра «Из истории анализа данных в России».
Начал выступление Борис Миркин с рассказа о своей научной карьере, которая включала работу в России (Саратов, Новосибирск, Москва), во Франции, США, Германии и Великобритании, затем подробно остановился на биографиях других замечательных советских и российских ученых.
«Проработав более 20 лет на Западе, я понял, что Россию там ценят только как страну с богатейшей культурой, но ее вклад в мировую науку не столь существенен», — отметил он. По мнению Миркина, это связано с двумя факторами: «во-первых, помешала политика «железного занавеса» — в наших библиотеках было тяжело найти иностранные книги, а переводились они лишь через несколько лет после публикации. А уж о прямых контактах с иностранными учеными и говорить не приходилось. Все это очень сильно тормозило развитие науки в СССР». В качестве второго фактора Борис Миркин выделил, то, что неоспоримым научным локомотивом в то время были США, и большинство научных разработок совершались именно в этой стране, которая являлась стратегическим соперником СССР на мировой арене.
«Россия была исключена из мирового научного процесса, однако в какой-то мере это позволило выработать свой оригинальный подход к решению многих задач, например, в сфере математической статистики были выдвинуты и доказаны многие крупные теории», — отметил Миркин. В связи с этим, в первую очередь, стоит выделить теорию Бомгарта, которая вошла в сокровищницу мировой науки, правда, в большей степени она нашла свое применение не в сфере анализа данных, а в сфере психологии. Достаточно подробно Борис Миркин остановился на судьбе незаурядного советского ученого Чеснокова, создателя так называемой социологической линейки. Чесноков также занимался проблемами совмещения изображений в фас и в профиль., и в этой области стал настоящим новатором. К сожалению, он так и не смог защитить докторскую диссертацию, и это сильно затормозило его исследования, а спустя много лет этой проблемой занялись совершенно другие люди.
Борис Миркин назвал имена и других ученых, сделавших большой вклад в науку анализа данных, среди них такие известные люди, как Фельд, ставший автором теории о сравнении объектов по их подмножествам, организатор международных научных форумов Айвазян, Дробышев, внесший неоценимый вклад в разработку матрицы связей между объектами, и многих других.
В завершение своего выступления Борис Миркин поделился житейской мудростью, полезной для начинающих ученых, а впрочем, и для людей далеких от науки. Например: как правильно выбрать направление своей деятельности? «Это в первую очередь зависит от склада ума, — считает Миркин. — В самом общем виде можно выделить две группы: индуктивную и дедуктивную. То есть предпочитаете ли вы двигаться от частного к общему, или от общего частному. Именно в зависимости от этих ваших склонностей и стоит выбирать направление своей деятельности». Кроме того, по мнению Бориса Миркина, очень важно знать свои общие склонности и таланты. В частности, чем предпочитаете заниматься — аналитической работой или, скажем, административной.
Но проблемы возникают и в процессе выбранной работы. Для того чтобы спокойно относиться к неудачам, нужно помнить о нескольких простых вещах. «Во-первых, следует правильно осознавать масштабы своей деятельности и свою роль в мировой науке, — сказал Миркин, — а во-вторых, нужно уметь правильно к себе относиться: с одной стороны, не слишком серьезно, но и не впадать в самоуничижение». А если что-то не получается, то всегда можно попробовать свои силы в смежной области.
Отдельно Борис Миркин остановился на правилах публикации научных работ. По его мнению, здесь главное уметь вписаться в существующий научный процесс, а не пытаться его изменить. Споры с бюрократами от науки — бесполезное дело. Ученый должен уметь их обходить — «в конце концов, все изобретения были сделаны вопреки желаниям этих самых бюрократов», резюмировал Борис Миркин.

Мои комментарии к заметке:

Мне кажется, всегда следует благодарить тех, кто про вас что-то пишет, что я и делаю. Вместе с тем имеется несколько существенных неточностей. Вот они.

1. *Я не говорил, что большинство разработок совершалось в США. Я говорил, что научная разработка оставалась безвестной, если она не была конституирована в США.*
2. *Главное российское достижение – теория VC-complexity, SVM classifier (метод опорных векторов) and kernel functions (ядерные функции, потенциальные функции) – пропущено (Вапник и Червоненкис - теория сложности, Вапник и Лернер – метод опорных векторов, Мучник, Айзерман, Браверман и Розоноэр – потенциальные функции).*
3. *Чесноков и Файн, два очень разных человека, вероятно, никогда не встречавшиеся, объединены.*
4. *«Индуктивное-дедуктивное» я не упоминал. Мне это кажется несущественным, да и скорее всего несуществующим на уровне когнитивной системы. Возможно, я мог рассказать примеры*

индивидуальных когнитивных систем: одни ориентированы на общие представления, другие – на частные случаи. (Но они так и сидят где сидят.)