

Синдромный анализ: новые вызовы¹

В.С. Переверзев-Орлов, В.Г. Трунов

Институт проблем передачи информации имени А.А. Харкевича РАН, Москва, Россия

Поступила в редколлегию 17.11.2008

Аннотация—Синдромный анализ – любопытный вид анализа данных, возникший при попытках формализации, развития, порождения и распространения знаний в медицине. Синдромный анализ – построение синдромных сетей при анализе баз прецедентов под контролем (или без) уже существующего профессионального знания. В результате такого анализа синтезируется не заданной заранее сложности пороговая многоуровневая сеть – разновидность нейронных сетей или многоуровневых персептронов, напоминающая байесовские сети, но свободная от их недостатков.

1. ВВЕДЕНИЕ

Первая фаза Синдромного Анализа (**СА**), начавшаяся для нас в конце семидесятых с шока от неожиданной силы примитивного голосующего порогового формализма в решении весьма нетривиальной задачи выявления настороженности в применении сильнодействующей терпии при остром инфаркте миокарда [1], приведшая к углубленной экспериментальной разработке этого подхода, завершилась установлением соответствия его синдромно-симптомной парадигме врачебного мышления, приведшего – в свою очередь – к попыткам создания на этой основе профессиональных баз знаний для Партнерских систем в медицине и разработке подходов к решению на этой основе задач развития и порождения знаний, а также – передачи знаний в новом типе исследовательско-игровых обучающих систем. Для поддержки всех этих возможностей были созданы в нескольких модификациях программы типов **Sand** и **Sigm** для автоматического и интерактивного синтеза синдромных сетей, в которых отрабатывались возможности новых интерфейсов в виде мозаичного коврика и гистограммного дерева, оказавшихся удобными и весьма эффективными средствами организации взаимодействия исследователя с базой данных в процессе направленного поиска и верификации знаний под контролем уже существующих знаний [2–7].

Все эти исследования показали, что синдромные пороговые сети являются крайне интересным объектом, в котором роль структуры сети и порогов в суммирующих симптомах – бинарные входы – в узлах оказываются принципиально важными факторами, требующими самостоятельного понимания. К сожалению, подходящей математики для этого пока, кажется, не существует, поэтому приходилось обходиться экспериментальными исследованиями и качественным их анализом, показавшими, что на основе синдромных сетей можно заниматься эффективным представлением понятий [8] и иметь возможность выхода через это на решение смысловых проблем языка. Но все это по большому счету пока еще ждет своего часа.

2. МОДЕЛИРОВАНИЕ ПРОЦЕССОВ

Круг возможностей **СА** этим далеко не исчерпывался. Так, в какой-то момент стало ясно, что на основе **СА** можно заниматься не только поиском эффективных решающих правил

¹ Работа поддержана Грантом Президиума Российской академии наук по программе “Интеллектуальные компьютерные системы”, проект № 2.3.4: “Партнерская система как ядро Усилителя интеллекта специалиста”, 2006–2008 гг.

для решения задач обучения распознаванию многомерных объектов, формализации понятий и развития знаний, но что также этот подход может быть применен и для анализа динамических процессов, рассматриваемых как многомерные векторные функции запаздывающих аргументов. Намек на такого рода возможности содержался все в той же медицине, где врачи выработали на протяжении тысячелетий не только синдромно-симптомный взгляд на состояние пациента, но и нашли способ свертывания предшествующей контакту истории заболевания в множество обобщающих оценок-функционалов, наиболее ярко представляемых в так называемом анамнезе. Такого же рода обобщающие оценки широко используются и при регулярном наблюдении за пациентом, например, в стационаре.

Итак, стало ясно, что синдромность и анализ динамических процессов не противопоставляются друг другу, а – напротив – естественно сочетаются. Это позволило определить новую фазу разработки и исследований **СА**, относящуюся к динамическим процессам, частным случаем которых можно было бы считать все те задачи и исследования, с которыми мы имели дело до того. С этим пониманием пришло и понимание того, что нужно искать подходящие для такого исследования задачи.

Очевидно, что динамических процессов, которые можно было бы исследовать, – великое множество, однако предыдущий опыт однозначно подсказывал, что задача должна удовлетворять некому набору требований, несоблюдение которых, могло бы сильно затруднить продвижение к желаемой цели. К таким требованиям – с учетом специфики нашего прежнего опыта – в первую очередь относятся:

- легкость доступа к необходимым данным
- наличие специалиста, готового участвовать в работе такого рода
- достаточная простота выбираемой конкретной задачи при условии, что все основные свойства общей решаемой задачи ей присущи
- возможность применения методов обучения распознаванию для поиска решений.

Последовательность этих требований не связана с их важностью, но последнее стоит несколько особняком, поскольку такого рода методы в явном виде обладают потенциалом развития знаний. Перебирая разные варианты задач в свете такого рода требований, мы достаточно быстро вышли на задачи биржевого предсказания. С ними по-началу, естественно, возникли некоторые проблемы переходного периода, но их в какой-то мере за достаточно короткое время удалось преодолеть, и у нас возникла реальная возможность работы с динамическими процессами, где реальная сложность рассматриваемых задач, не считая проблем содержательной их постановки, определялась по существу уже только нашими собственными возможностями. Таким образом, появилась возможность начать формулировать требования к инструментальной системе поддержки исследований.

Требования формируются несколькими обстоятельствами:

- было ясно, что задачи, которые предстояло решать с помощью создаваемой технологии, по своей сложности в общем случае не должны уступать задачам, возникающим в медицине в связи с формированием диагностических гипотез и т.п. по наблюдаемым характеристикам исследуемого объекта (пациента)
- было ясно, что динамические процессы в системах интересующей сложности, как и в медицине, должны характеризоваться по возможности многими своими внешними проявлениями
- было ясно, что наблюдаемые проявления таких процессов по существу могут быть лишь вырожденными отображениями того, что на самом деле выражает суть этих процессов, но не доступно непосредственно внешнему наблюдателю

- было ясно, что поскольку внешние проявления – вырожденные отображения непосредственно не наблюдаемого, то желательно иметь как можно больше различных по характеру такого рода отображений
- было ясно, что предсказание поведения процесса можно строить по множеству его текущих функционалов, в разной степени, учитывающих историческое в процессе
- было ясно, что возможная избыточность совокупности доступных наблюдению вырожденных описаний не должна рассматриваться как препятствие на пути к достижению цели, лишь бы в этих описаниях содержалось нужное для решения
- было ясно, что с такого рода избыточностью можно будет справиться, используя адекватные алгоритмы обучения в качестве инструмента моделирования законов, порождающих изучаемый процесс
- было ясно, что по возможности следует использовать дополнительные знания относительно изучаемых процессов
- наконец, было ясно, что алгоритмам обучения распознаванию присуща и такая способность как направленный синтез фильтров-функционалов, максимально соответствующих целям и критериям решаемых задач такого рода.

По совокупности всех этих соображений и требований было решено, что наилучшим в качестве источника для постановки задач, связанных с динамическими процессами, является какой-нибудь сегмент биржи, достаточно автономный, чтобы в первом приближении его можно было бы рассматривать в качестве самодостаточной системы, с достаточно сложными внутренними связями и обеспечивающий желательные полноту и доступность своих внешних проявлений. После некоторых консультаций с понимающими в этом коллегами было решено остановиться на валютном рынке, ограничившись задачами предсказания поведения цен ведущих валют и тесно связанных с ними металлов, всего - порядка десятка таких инструментов.

Уточненный анализ возникающих при этом задач показал, что можно рассматривать исходные ценовые потоки, порождаемые этими инструментами по крайней мере с двух точек зрения:

- С точки зрения текущего представления информации о них с помощью множества функционалов, известных в “биржевом техническом анализе” как индикаторы. Специалисты считают, что реально хоть сколько-нибудь существенную информацию о процессе несут примерно сто-двести таких индикаторов.
- Но можно подойти и с более активной точки зрения и рассматривать активный синтез моделей процессов, используя для этого присущие методам обучения распознаванию возможности. В этом случае придется ориентироваться не только на текущие значения цен, но и на их величины в течение некоторой истории, предшествующей текущим значениям этих цен. По мнению специалистов, длина интервала предыстории должна по крайней мере в сто-двести раз превосходить тот элементарный квант времени – бар, величина которого определяет временной масштаб событий в исследуемом процессе.
- Наконец, утверждалось, что интервалы времени, охватывающие порядка 5–7 тысяч исследуемых тактов-баров могут считаться вполне достаточными для того, чтобы выявить на них основные тенденции поведения изучаемых процессов.

Таким образом, оказалось, что если мы хотим иметь инструмент для исследования такого рода процессов, то он должен обладать возможностью обеспечения проведения исследований в режиме, близком к реальному времени, с многомерными векторными функциями, заданными примерно 5–7 тысячами срезов и с размерностью порядка 4000. Это – так сказать, внешние параметры. Но существенное развитие в этом новом инструменте получили и его внутренние возможности. В первую очередь это относится к тому, что:

- Используется функция гигиены анализируемых данных. В этой фазе анализируются свойства различных переменных (столбцов), что позволяет определить, является ли эта переменная числовой, ранговой или категориальной. Такие гипотезы системы могут корректироваться исследователем, обладающим большей полнотой информации. Также определяются диапазоны разброса значений таких переменных и их основные статистики, обнаруживаются неинформативные переменные и пропуски в данных.
- Для ускорения анализа предусмотрена возможность использования различных вариантов шкалирования первичных переменных. Можно использовать как равномерное, так и квантильное квантование, привязанное к общему распределению объектов в выборке, к распределению, скорректированному разным представительство классов вплоть до ориентации на взвешенную комбинацию определенных классов.
- Используются несколько вариантов визуализации распределений данных, включая множественные двумерные гистограммы с объектами, расцвеченными классами.
- Наиболее существенному развитию подверглась собственно функция синтеза синдромов как в интерактивном, так и в автоматическом режимах. Главным здесь оказалось использование разнообразных критериев для управления процессом синтеза на уровне подбора претендентов на включение в создаваемый синдром и гибкое управление теми подмножествами переменных (как исходных, так и синтезированных), из которых будут выбираться претенденты на включение в синдром.
- Наконец, отработана возможность направленного синтеза желаемого синдрома в процессе конструирования его структуры и оптимизации ее и порогов на основании статистик данных в узлах создаваемой структуры.

3. ЗАКЛЮЧЕНИЕ

В настоящее время ведется интенсивное исследование созданного инструмента на реальных данных. Наши проблемы существенно осложняются отсутствием участия в работе биржевых специалистов, которые могли бы ставить действительно интересные в этой области задачи. Поэтому пока мы вынуждены ставить такие задачи самостоятельно, исходя из знаний, почерпнутых из соответствующих книг, чего явно недостаточно для серьезной профессиональной работы, как, например, это было при решении медицинских задач. И хотя работа далеко не завершена, уже получены вполне обнадеживающие результаты, показывающие, что динамическое порождение синдромных моделей поведения цен и их динамическая адаптация к условиям меняющихся рынков могут быть основой для решения такого рода проблем предсказания.

СПИСОК ЛИТЕРАТУРЫ

1. Алексеевская М., Переверзев-Орлов В., Саблин В.. Прогнозирование врачебной тактики. *Вопросы кибернетики*. 1979. Вып. 49.
2. Pereverzev-Orlov V.S., Stenina I.I., Trunov V.G.. Syndrome Analysis of Data. *Pattern Recogn. and Image Anal.*, 1993, Vol. 3, No. 4, pp. 500–507.
3. Ващенко Е.А., Гуров Н.Д., Лайков Т.В., Переверзев-Орлов В.С., Стенина И.И. Интерактивные средства создания и анализа деревьев решений, *2-ая Всероссийская с участием стран СНГ конференция "Распознавание образов и анализ изображений: новые информационные технологии"*, Ульяновск, 1995.
4. Vashchenko E., Gurov N., Laikov T., Pereverzev-Orlov V., Stenina I. Interactive Tools for Creation and Analysis of Decision Trees, *Pattern Recognition and Image Analysis*, 1996, vol. 6, no. 2, pp.367–368.
5. Vitushko M.A. A Support System for Knowledge Base Design. *Pattern Recognition and Image Analysis*, 1997, vol. 7, no. 2, pp. 266–275.

6. Vashchenko E.A., Vitushko M.A., Gurov N.D., Pereverzev-Orlov V.S., Stenina I.I. Knowledge and Data Cooperation. *Pattern Recognition and Image Analysis*, 1998, vol. 8, no. 2, pp. 25–41.
7. Витушко М.А., Гуров Н.Д., Переверзев-Орлов В.С. Синдромное прогнозирование изменчивости. *Сб. докладов 10-й Всероссийской конференции “Математические методы распознавания образов” (ММРО-10)*, Звенигород, 2001, с. 28–30.
8. 8. Vitushko M., Gurov N., and Pereverzev-Orlov V.. A Syndrome As a Tool for Presenting Concepts, *Pattern Recognition and Image Analysis*, 2002, vol. 12, no. 2, pp. 194–202.