SYBASE®
An SAP Company

ИНФОРМАЦИОННЫЙ ДОКУМЕНТ

PMML: ускорение внедрения предсказательной аналитики в эпоху «больших данных»

Без адекватных средств моделирования и анализа «большие данные» абсолютно бесполезны. Своей эффективностью предсказательная аналитика в значительной мере обязана сложным математическим моделям, применяемым для анализа огромных информационных массивов с целью извлечения полезных сведений. Для создания предсказательных моделей необходимы известный опыт и квалификация. А перенос моделей в базы данных требует еще большего искусства и временных затрат.

В сущности, значительные затраты времени на имплементацию моделей и стали причиной разработки индустрией бизнес-аналитических средств языка PMML (Predictive Model Markup Language). Его применение позволяет существенно ускорить получение результатов предсказательного моделирования.

Sybase IQ — наиболее многофункциональное и самое быстродействующее специализированное средство бизнес-аналитики, выполняющее счет по сложным предсказательным моделям. Занимая ведущие позиции в мире по числу инсталляций среди поколоночных СУБД, Sybase IQ предлагает непревзойденные возможности предприятиям, внедряющих предсказательную аналитику для информационного обеспечения своей деятельности — начиная от выявления мошенничества в реальном времени и оперативных онлайновых промо-акций и кончая скоростной обработкой больших объемов транзакций и подготовкой углубленных стратегических отчетов для советов директоров. Эти и многие другие бизнес-приложения используют уникальное свойство Sybase IQ во многих случаях рассчитывать предсказательные аналитические модели буквально в мгновение ока. (См. стр. 5, «Разделение всех ресурсов: преимущество для моделировщиков».)

При этом максимального быстродействия приложений предсказательной аналитики в Sybase IQ удается достичь тогда, когда модели импортируются в базу данных. Вплоть до настоящего времени имело место большое запаздывание между моментами завершения работы над аналитической моделью (которую выполнял разработчик приложений) и ее помещения в рабочую среду. Это запаздывание приводит к расходованию ресурсов, при этом исполнение части управленческих решений откладывается в ожидании того, пока технические специалисты воссоздадут модель в корпоративном хранилище данных — в итоге решения теряют актуальность.

Реализовав в своем продукте PMML, Sybase устранила одно из основных препятствий на пути внедрения аналитических приложений, способных изменить ход игры. Отныне в распоряжении пользователей — одно из самых быстродействующих автоматизированных средств для импорта моделей в Sybase IQ.

Центральная логика этих приложений использует аналитические модели, в основе которых — выверенные алгоритмы и точные математические расчеты. Предсказательная аналитика слишком сложна, чтобы ее можно было реализовать в форме простых SQL-запросов по отношению к хранилищу данных. Разработка эффективных алгоритмов информационной проходки и статистических моделей для предсказательной аналитики требует работы талантливых математиков и опытных исследователей, применяющих передовые средства моделирования.

После того как разработчик завершит работу над моделью, ее необходимо адаптировать к рабочей базе данных, отличной от среды, в которой модель строилась. Это отличие порождает серьезную проблему. Традиционно готовая модель переписывалась с самого начала для целевой базы данных, при этом сами данные также преобразовывались в форму, пригодную для использования. Возможности разных ИТ-департаментов неодинаковы, поэтому указанный процесс мог отнимать недели, месяцы, а в иных случаях даже год или более.

Понятно, что в течение столь длительного срока требования к модели могут существенно измениться. В результате окончательно готовые модели оказываются нерелевантными. Хуже того, в ожидании, пока модель будет готова, можно упустить ценные возможности, открывающиеся на рынке.

Язык PMML — это отраслевой стандарт, обеспечивающий быстрый экспорт и импорт сложных аналитических моделей для совместимых приложений моделирования. Во многих случаях он позволяет срок переноса модели в рабочую среду с нескольких недель до нескольких часов.

PMML: СТАНДАРТНЫЙ ЯЗЫК МОДЕЛИРОВАНИЯ

Язык PMML (основан на XML) разработан Data Mining Group — отраслевым консорциумом, в состав которого входят компании IBM/SPSS, MicroStrategy, SAS и ряд других. В разработке стандарта приняли участие такие авторитетные фирмы, как Microsoft, SAP, BusinessObjects и Tibco. Стандарт хорошо продуман и за время 10-летнего существования вполне доказал свою эффективность. По состоянию на декабрь 2011 года используется его 4-я версия.[1]

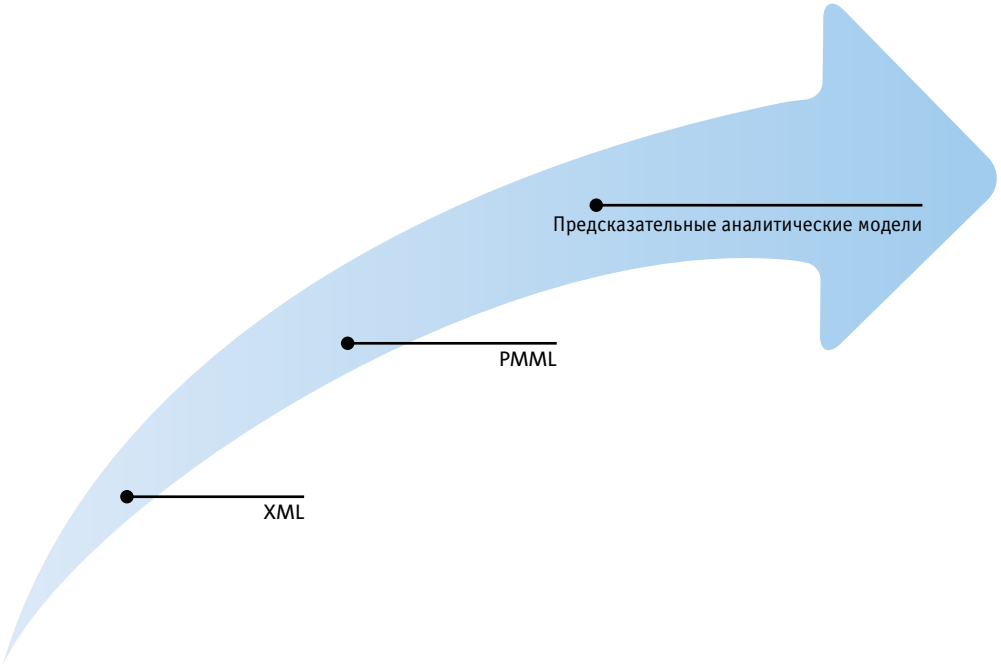

Рис. 1. Язык PMML, основанный на языке XML, являющимся отраслевым стандартом, будет способствовать интенсификации внедрения предсказательного моделирования в самых разных областях.

Будучи фактическим стандартом для представления статистических моделей и схем информационной проходки, язык PMML сокращает длительный, требующий многократных итераций процесс разработки, в котором аналитические проекты на этапе движения от создания модели до исполнения в рабочей БД подчас буквально увязают. Благодаря автоматизации перекодировки модели для обеспечения ее соответствия среде исполнения сроки имплементации модели существенно сокращаются. Еще одна причина ускорения времени разработки кроется в том, что благодаря широкому внедрению PMML в индустрии аналитики могут продолжать использовать привычные средства моделирования.

Стандарт PMML — мощный и всеохватывающий.[2] Документы в этом формате состоят из следующих разделов. **Заголовок** (header) описывает сам документ и в частности содержит номер используемой версии стандарта. **Словарь данных** (data dictionary) определяет поля, используемые в предсказательной модели, а **схема проходки** (mining scheme) задает поля, присущие конкретной модели. Задаются также **таксономия** (taxonomy) модели, ее **статистика** (statistics), **цели** (targets), **вывод** (output) и **функции** (functions). Кроме того, предусмотрены различные **преобразования данных** (data transformations), такие как нормализация, агрегация и отображение значений. И, конечно, в документе присутствует сама **модель** (model), содержащая такие атрибуты, как имя, функция, алгоритм и др.

Стандарт PMML приспособлен к обработке больших массивов данных. С помощью сочетания арифметических и логических операций он позволяет задавать относительно сложные предварительные процедуры обработки, обеспечивающие приведение данных на этапе разработки к форматам, требуемым средой исполнения. В эпоху «больших данных» это весомое преимущество. Данные, используемые в ходе создания модели, должны быть настолько полны, чтобы построенная на их базе модель могла обрабатывать реальные данные во всем их объеме с высокой эффективностью.

Процедура экспорта модели в формате PMML из любого средства моделирования, которые в изобилии представлены на рынке, проста. После проверки модели в среде разработки для ее вывода в PMML используется встроенный пошаговый механизм. И хотя в разных средствах этот механизм различен, как правило, вызывать функцию экспорта PMML не сложнее, чем сохранить файл в заданном формате с помощью диалога «Save As…».

Рис. 2. Разработчик может использовать любое средство моделирования и экспортировать готовые модели в формате PMML. Для запуска модели в Sybase IQ используется подключаемый модуль Sybase Universal PMML Plug-in.

Поддержка PMML в Sybase IQ реализована посредством подключаемого модуля Sybase Universal PMML Plug-in. Он автоматически распознает совместимые модели, при этом поддерживаются все версии стандарта PMML начиная с 2.0. Данный программный модуль был разработан компанией Zementis Inc., лидером сообщества PMML.[3] Модуль обеспечивает многообразие возможностей моделирования (см. «Возможности моделирования»). Аналитики могут создавать модели самых разных видов, используя разнообразные средства разработки, и без труда экспортировать их в рабочую БД Sybase IQ — при этом результаты будут готовы практически немедленно.

Использование PMML дает такую скорость имплементации моделей, что по мере изменения условий деятельности их можно обновлять и быстро вновь вводить в эксплуатацию. Например, после слияния двух компаний розничной торговли аналитическую модель, используемую менеджментом для оценки общего объема продаж, можно быстро настроить так, чтобы прогнозы строились с учетом деятельности магазинов новой компании, охватывающих участки тех же районов, что и прежние магазины.

Возможности моделирования

Используя Universal PMML Plug-in для Sybase IQ, аналитики могут строить самые разнообразные широкое предсказательные аналитические модели для быстрого скоринга. В их числе:

- деревья решений для классификации и регрессии;
- нейросетевые модели: обратное распространение, радиальные базисные функции и «нейронный газ»;
- векторные машины для регрессии, двоичной и мультиклассовой классификации;
- линейная логистическая регрессия (двоичная и полиномиальная);
- простые байесовские классификаторы;
- общие и генерализованные линейные модели;
- регрессия Кокса;
- модели с набором правил (плоские деревья решений);
- кластерные модели: распределение, центральное распределение, двухшаговая кластеризация;
- карты балльных оценок (начисление баллов по категориальным, непрерывным и сложным атрибутам);
- правила ассоциаций;
- множественные модели (композиция, ансамбли моделей и сегментация).

Модуль также позволяет определять словари данных, обрабатывать отсутствующие и недопустимые значения, а также выполнять предварительную обработку данных.

PMML обеспечивает простое использование предсказательных аналитических моделей несколькими приложениями одновременно. Так, можно обучить модель в системе, применяемой для планирования выхода годного продукта с линии сборки, выразить ее в PMML, настроить, оттестировать в среде разработки, а затем быстро перенести в другую систему, где использовать, например, для предсказания выработки продукта в другом производственном процессе.

Кроме того, PMML — исключительно гибкий стандарт, рассчитанный на удовлетворение потребностей современных специалистов — бизнес-аналитиков, применяющих передовые методы моделирования. Например, поскольку предсказательные модели рассчитаны на решение конкретных проблем и их преимущества проявляются только при применении по прямому назначению, в сложных случаях одной модели для поиска решения будет недостаточно. PMML позволяет строить приложения с несколькими моделями, включая ансамбли моделей. Каждую модель можно экспортировать в PMML и выполнить в рабочей среде так, как запланировано в приложении.

Еще одно преимущество использования PMML обусловлено проблемой текучести кадров. В некоторых случаях отсутствие возможности выразить модель в доступной для посторонних форме с использованием общепринятого и понятного стандарта приводит к тому, что когда разработчик покидает организацию, информация о том, как модель функционирует, исчезает вместе с ним. Может оказаться так, что модель нельзя будет приспособить к изменившимся условиям. Применение PMML позволяет сохранить накопленные знания на будущее.

РАЗДЕЛЕНИЕ ВСЕХ РЕСУРСОВ: ПРЕИМУЩЕСТВО ДЛЯ МОДЕЛИРОВЩИКОВ

Архитектура Sybase IQ дает этому продукту хорошо выраженное преимущество над другими средствами при счете по аналитическим моделям. Использование уникальной системы массово-параллельной поколоночной обработки с разделением всех ресурсов PlexQ™ обеспечивает отсутствие для аналитических моделей каких-либо ограничений. Вычислительные ресурсы, память, а также дисковое пространство систем могут наращиваться в соответствии со сколь угодно высокими требованиями.

Таким образом, разработчикам моделей нет необходимости учитывать ресурсные ограничения при создании PMML-совместимых приложений предсказательной аналитики. В отличие от архитектур без разделения ресурсов, в PlexQ применен подход совместного использования всех ресурсов, когда рабочая нагрузка по обработке запросов динамически распределяется между всеми узлами, составляющими кластер PlexQ. Автоматический балансировщик нагрузки PlexQ постоянно перераспределяет ее таким образом, чтобы избежать конкуренции пользователей за системные ресурсы, тем самым обеспечивая высокую и предсказуемую производительность и эффективность использования ресурсов для широкого спектра одновременных рабочих нагрузок.

В основе технологии PlexQ — принцип распределенной обработки запросов (Distributed Query Processing — DQP). Он заключается в разбиении запроса на части и поручении обработки этих частей нескольким серверам Sybase IQ, составляющим кластер. Поскольку обработка выполняется одновременно, время исполнения запроса уменьшается. Распределенная обработка в PlexQ организована таким образом, что пространство хранения можно наращивать независимо от вычислительных мощностей (которые определяются количеством одновременно работающих пользователей и их требованиями к системе), удовлетворяя параметрам соглашений об уровне обслуживания.

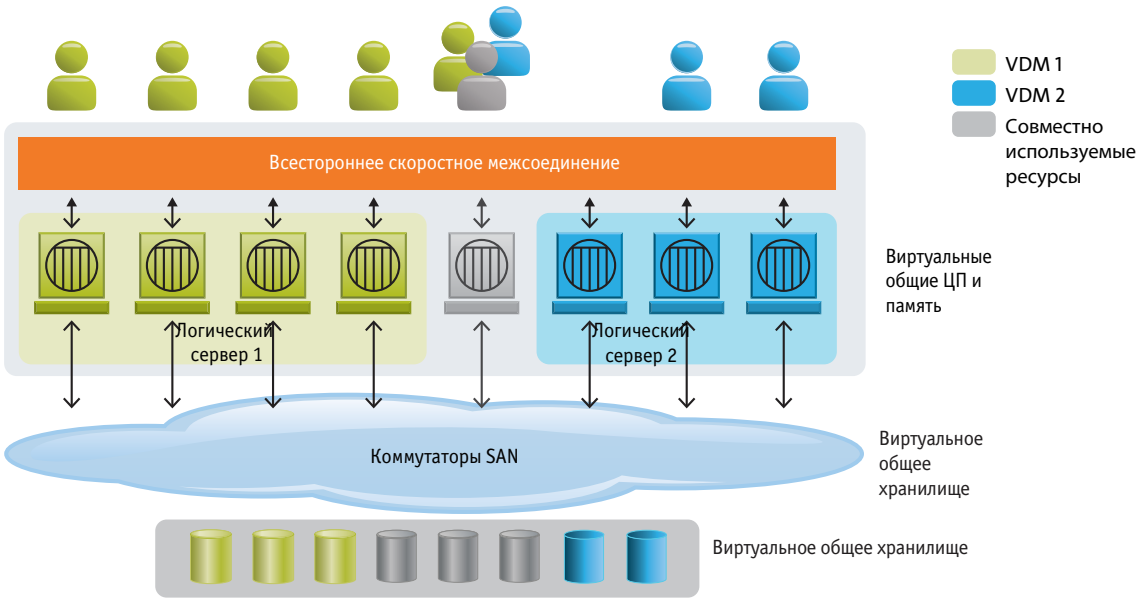

Рис. 3. Реализованная в Sybase IQ архитектура массово-параллельной обработки с разделением всех ресурсов PlexQ обеспечивает масштабирование для любых предсказательных аналитических моделей на PMML.

ПРИМЕР ИСПОЛЬЗОВАНИЯ: ЗДРАВООХРАНЕНИЕ

PMML позволяет ускорить внедрение аналитических моделей в любых отраслях — будь то связь, производство, финансовые услуги или розничная торговля. Эти модели обеспечивают необходимой информацией практически все подразделения на всех уровнях управления. Взять, к примеру, отрасль здравоохранения: преимущества стандарта PMML в данном случае дают гораздо более, нежели просто эффективность: ускоряя внедрение аналитических средств в процесс принятия управленческих решений, стандарт способствует более быстрому решению вопросов буквально жизни и смерти.

Как было отмечено выше, ПО Sybase поддерживает широкое разнообразие моделей в формате PMML, используемых в приложениях предсказательной аналитики. Один из примеров моделей, доказавших свою эффективность в области здравоохранения (как исследований, так и практической медицины) — искусственные нейронные сети.[4] Объединяя данные, поступающие от кардиоваскулярных датчиков, они дают цельное представление о состоянии кардиоваскулярной системы. Кроме того, искусственные нейронные сети — превосходное средство для быстрого анализа снимков с целью выявления опухолей и других патологий.[5]

Архитектура с разделением всех ресурсов имеет существенное значение для организаций, стратегическая задача которых — поддерживать централизованный репозиторий данных, служащий единым источником согласованной и непротиворечивой информации. Благодаря отсутствию ограничений в масштабировании, свойственных другим архитектурам, технология PlexQ позволяет аналитикам любого подразделения организации строить сложные, требовательные к ресурсам модели, которые, задействуя большие массивы данных общего пользования, могут использоваться сколь угодно интенсивно и каким угодно количеством пользователей. Большие массивы данных служат источником для моделирования в самых разных областях деятельности, в частности в здравоохранении.

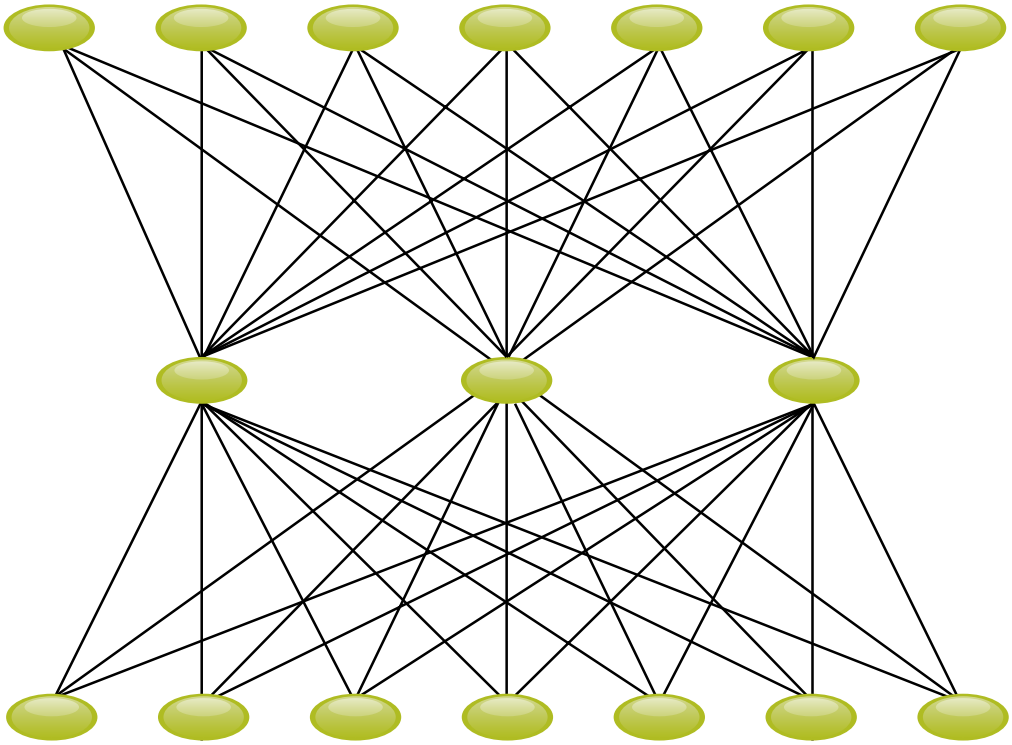

Рис. 4. Простая искусственная нейронная сеть: наверху входные нейроны, в середине скрытые нейроны, внизу выходные нейроны.

Возможно также решение задач предсказательной аналитики в реальном времени. В частности, врачи и медсестры отделений интенсивной терапии уже применяют модели предсказательной аналитики для контроля показателей жизнедеятельности недоношенных новорожденных, поступающих от многочисленных датчиков. Аналитическое ПО может выявить отклонения от нормы прежде, чем медицинский работник что-нибудь заметит — тем более, что ему, как правило, доступны лишь данные на определенный момент времени, а не в динамике.[6]

В той же мере аналитические модели полезны и администрациям учреждений здравоохранения. Так, закон США о защите пациентов и доступном медицинском обслуживании 2010 года требует от учреждений, действующих по программам Medicare и Medicaid, ввести новые правила повторного лечения пациентов, предусматривающие дополнительное амбулаторное лечение в определенных случаях и сокращающие объем возмещения больницам в некоторых случаях повторной госпитализации.[7] Согласно отчету, опубликованному в New England Journal of Medicine, незапланированные повторные госпитализации обходятся для программы Medicare более чем в 17 млрд. долларов в год. Новые правила вступают в силу в октябре 2012 года. В идеале больницы должны оценить, какие группы пациентов подпадают под новые правила повторной госпитализации и скорректировать лечение таким образом, чтобы минимизировать потребность в последующей госпитализации. Разработка и быстрый ввод в эксплуатацию приложений, которые, к примеру, помогли бы определить, каким категориям пациентов лучше всего подходит амбулаторное лечение, были бы весьма полезны больницам, стремящимся выполнить новые жесткие требования правительства в части повторной госпитализации.

Обнаружение мошенничества — еще одна область экономики здравоохранения, где аналитические средства могут оказаться весьма полезными. По данным Федерального бюро расследований, убытки от воровства обходятся народному хозяйству США в 60 млрд. долларов ежегодно. Национальная ассоциация по предотвращению мошенничества в области здравоохранения (National Health-Care ANti-Fraud Association — NHCAA), которая оценивает объем мошенничества в области медицинского страхования примерно в 3% всех отраслевых расходов, отмечает: «Большая часть мошенничества в области медицинского страхования совершается крошечным количеством недобросовестных учреждений здравоохранения».[8] Это означает, что без помощи специальных программ люди не в состоянии выявить мошеннические операции среди множества законных. Мощные приложения предсказательной аналитики — идеальное средство выявления мошеннических страховых требований до того, как они будут оплачены.

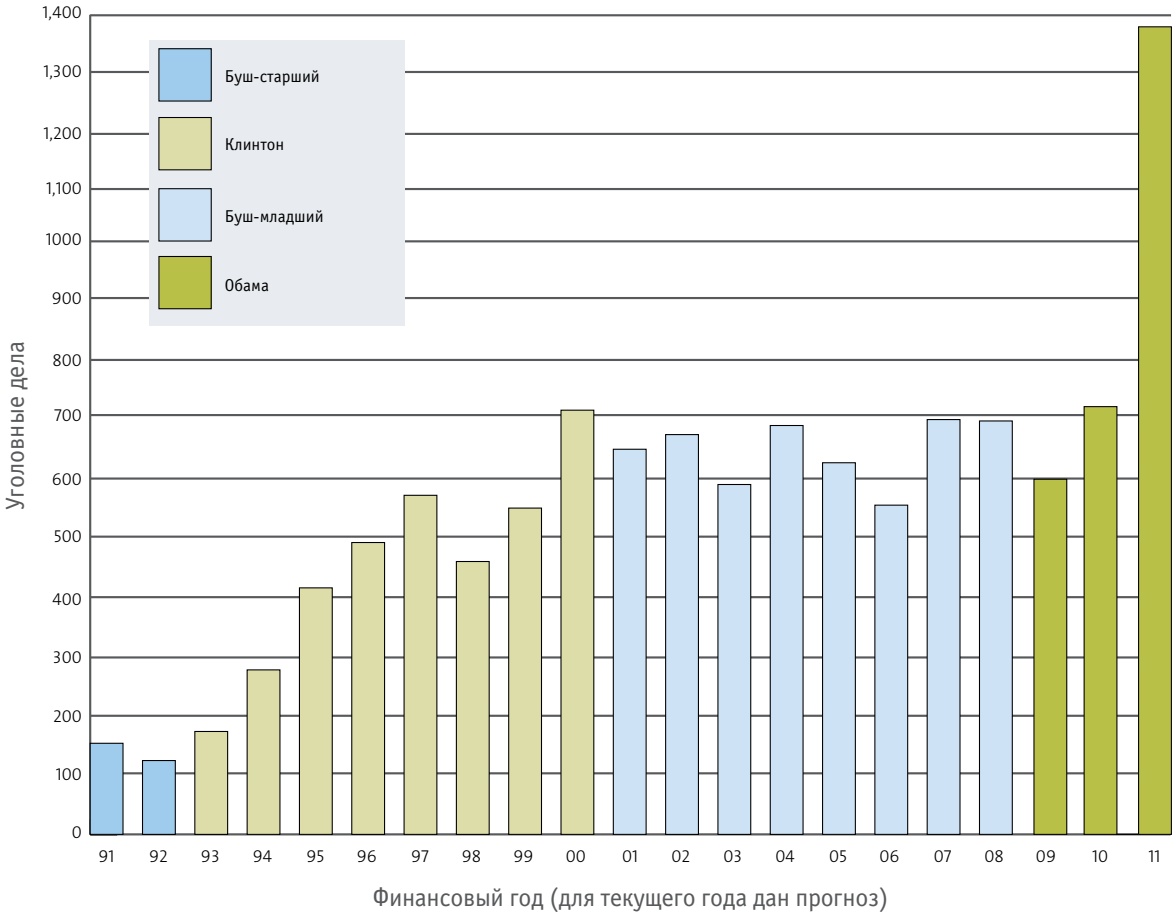

Рис. 5. Количество уголовных дел в связи с мошенничеством в здравоохранении, возбужденных федеральными властями США за последние 20 лет.

Как видно из рис. 5, правительство США начало активно преследовать мошенников в области здравоохранения, а также взыскивать с них средства для возмещения вреда, нанесенного страховым учреждениям, таким как Medicare. По данным NHCAA, каждые 2 млн. долл., направленные частными страховыми фирмами на борьбу с мошенничеством, позволяют вернуть 17 млн. долл. незаконно потраченных средств.[9]

Предсказательная аналитика находит применение и в самых экзотических приложениях. Газета New York Times сообщила, что американское правительство намерено начать экспериментальный анализ сообщений латиноамериканских пользователей Twitter на предмет выявления эпидемий, грозящих перерасти в пандемии.[10] Исследователи предполагают, что, используя в качестве исходных данных миллионы записей в сервисе микроблогов, они смогут построить аналитическую модель, которая будет выявлять факты и места локализации опасных инфекционных заболеваний. Медицинские эксперты надеются, что, располагая этой информацией, они смогут гасить вспышки инфекции до того, как те приобретут глобальный масштаб.

Во всех перечисленных и подобных им случаях скорость имплементации моделей играет решающую роль. Будь то выявление мошеннических страховых требований, ускоренный переход на новые правила госпитализации, лечение в отделениях интенсивной терапии или мониторинг распространения инфекционных заболеваний, время — самый важный фактор. Без PMML сроки адаптации модели к рабочей среде становятся неприемлемыми. В конце концов, в здравоохранении эффект измеряется не только деньгами — задержка может стоить жизни новорожденного или здоровья многих людей.

1 Data Mining Group (2011). PMML version 4.1, http://www.dmg.org/pmml-v4-1.html.

2 Predictive Model Markup Language. Wikipedia: The Free Encyclopedia. Wikimedia Foundation.

3 A. Guazzelli, W. Lin, T. Jena (2010). PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics. CreateSpace (имеется на Amazon.com — http://www.amazon.com/dp/1452858268).

4 Christos Stergiou and Dimitrios Sinanos, Neural Networks, http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report. html#Conclusion.

5 Harjit Singh, editor, Artificial Neural Networks in Medicine and Biology, Springer-Verlag, 2000.

6 Alex Guazzelli, Predictive Analytics in Healthcare, Nov. 2011. http://www.ibm.com/developerworks/industry/library/ind-PMML3/.

7 Robert Kocher, "Hospital Readmissions and the Affordable Care Act," Journal of the American Medical Association, October 2011. http://jama.ama-assn.org/content/306/16/1794.extract.

8 NHCAA, The Problem of Health Care Fraud, http://www.nhcaa.org/eweb/DynamicPage.aspx?webcode=anti_fraud_resource_centr&wpscode=TheProblemOfHCFraud.

9 Coalition Against Insurance Fraud, Go Figure: Fraud Data, http://www.insurancefraud.org/healthinsurance.htm.

10 John Markoff, "Government Aims to Build Data Eye in the Sky," New York Times, October 10, 2011. http://www.nytimes.com/2011/10/11/science/11predict.html?pagewanted=all

Sybase, Inc.
Worldwide Headquarters
One Sybase Drive
Dublin, CA 94568-7902
U.S.A
1 800 8 sybase

www.sybase.com

Sybase CIS
115114, Москва,
Дербеневская набережная,
д. 7, стр. 16
+7 (495) 797-4774

www.sybase.ru

SYBASE®
An SAP Company