Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/

Том 8, №6 (2016) http://naukovedenie.ru/vol8-6.php

URL статьи: http://naukovedenie.ru/PDF/166TVN616.pdf

Статья опубликована 02.02.2017

Ссылка для цитирования этой статьи:
Пивоварова Н.В., Видунова С.И. Интеллектуальный анализ данных в фармацевтическом бизнесе // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 8, №6 (2016) http://naukovedenie.ru/PDF/166TVN616.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 004.623

Пивоварова Наталья Владимировна
ФГБОУ ВО «Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)», Россия, Москва
Доцент кафедры «Системы автоматизированного проектирования»
Кандидат технических наук
E-mail: pivovarova.natasha2013@yandex.ru

Видунова Светлана Игоревна[1]
ФГБОУ ВО «Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)», Россия, Москва
Бакалавр кафедры «Системы автоматизированного проектирования»
E-mail: Svetlana.Vidunova@gmail.com

Интеллектуальный анализ данных в фармацевтическом бизнесе

Аннотация. Рост объема данных в различных областях, а также необходимость их анализа для получения полезной информации приводит к тому, что многие аналитики сталкиваются с различными задачами. Сбор данных сам по себе не приводит к каким-либо результатам, необходимо рассматривать данные в качестве сырья для извлечения полезной информации. В данной статье рассматривается алгоритм Apriori для получения ассоциативных правил из множества накопленных данных аптек фармацевтического предприятия.

Ключевые слова: интеллектуальный анализ; аналитическая система; алгоритм Apriori; ассоциативные правила; фармацевтика; бизнес; прибыль

Введение

В результате быстрого развития современных компьютерных технологий, значительный прогресс достигнут в автоматизации ежедневной работы в офисе. Вследствие этого количество бизнес-данных, накопленных человечеством в электронном виде, растет быстрыми темпами. Огромное количество информации доступно для использования человеком, но большинство компаний не в состоянии в полной мере обработать и использовать такое количество сведений для стратегического планирования своего развития. В связи с этим стали активно развиваться новые технологии или методологии, которые имеют цель автоматически обнаруживать знания в базах данных (Knowledge Discovery in Databases), извлекать и отличать открытые знания от «сырых» бизнес-данных. Но прежде чем автоматизировать анализ данных, необходимо

[1] 105042, Москва, Измайловский проспект, д. 73А, кв. 309

провести интеллектуальный анализ данных или, говоря современным языком, использовать Data Mining.

Задачи Data Mining

Data Mining - это термин, используемый для описания процесса обнаружения скрытых закономерностей в базах данных, которые являются полезными в процессе принятия решений. Важное положение технологии Data Mining - нетривиальность разыскиваемых закономерностей [1]. Их поиск производится методами, не ограниченными рамками формального анализа. Поскольку конкуренция между фирмами во всех областях экономики растет, появляется необходимость использования новаторских методов захвата доли рынка. Различные фирмы, организации, предприятия собирают и анализируют информацию, которая затем перепродается другим компаниям для использования в своем бизнесе.

Классифицировать методы Data Mining можно по задачам, которые эти методы могут решать - задачи классификации, кластеризации, прогнозирования, либо разделить методы на две большие группы по принципу работы с исходными данными: непосредственное использования данных или сохранение данных, выявление и использование формализированных закономерностей [1]. Перечислим основные задачи Data Mining.

1) Классификация: метод ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks).
2) Кластеризация: обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена.
3) Прогнозирование и/или последовательность: методы математической статистики, нейронные сети и др.
4) Ассоциация: наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

Методы и алгоритмы Data Mining помогают фармацевтике не только получать прибыль, но и выявлять случаи мошенничества и злоупотреблений, а также получить помощь в принятии решений в области управления отношениями с клиентами [2, 7]. Например, обнаружение того, что два часто используемых препарата - антидепрессант пароксетин и правастатин, используемый для понижения уровня холестерина - увеличивают риск развития диабета, если они употребляются совместно. Было бы невозможно без метода ассоциаций и Data Mining в целом [3].

Data Mining в фармацевтическом бизнесе

Data Mining в фармацевтике может быть описан, в частности, как сбор информации, касающейся назначения лекарственных средств для конечного потребителя. Примером являются сбор данных продаж аптеки и их использования для определения шаблонов отдельных препаратов индивидуального назначения. Другие фармацевтические компании покупают эту информацию, так как она помогает им лучше нацеливать свои полевые силы (Полевые силы - сотрудники фармацевтических компаний, занимающиеся продвижением препаратов во врачебной аудитории и среди провизоров). Такая специфическая практика интеллектуального анализа данных получила широкое распространение, поскольку она показала большой потенциал повышения эффективности маркетинга и, в конечном счете, прибыли [4].

Типичным примером системы DM является специально разработанная аналитическая система [5] фармацевтической компании, которая решила проанализировать транзакции покупок всех аптечных сетей. Для нахождения закономерностей в покупках компания решила использовать ассоциативные правила, получаемые с помощью метода Apriori.

Для реализации метода необходимы конкретные данные, поэтому предварительно производилась детализация каждого чека каждой аптеки, входящей в сеть компании, несмотря на то, что данный способ хранения информации являлся дорогостоящим с точки зрения стоимости дискового пространства.

Компания использовала 27 рабочих станций UNIX для ежедневного интеллектуального анализа и 37 персональных компьютеров для передачи данных из 1230 аптек. Когда аптеки были закрыты, данные продаж, накопленные в течение дня, направлялись в штаб-квартиру через общую сеть для хранения в специальных файлах (рисунок 1). Для системы были специально разработаны 50 UNIXShell команд, которые в процессе конкретного запрошенного анализа объединялись в прикладную программу. Размер каждой полученной программы достаточно мал, что способствовало быстрому реагированию на различные запросы интеллектуального анализа данных и помогало справиться с большим количеством входных данных.

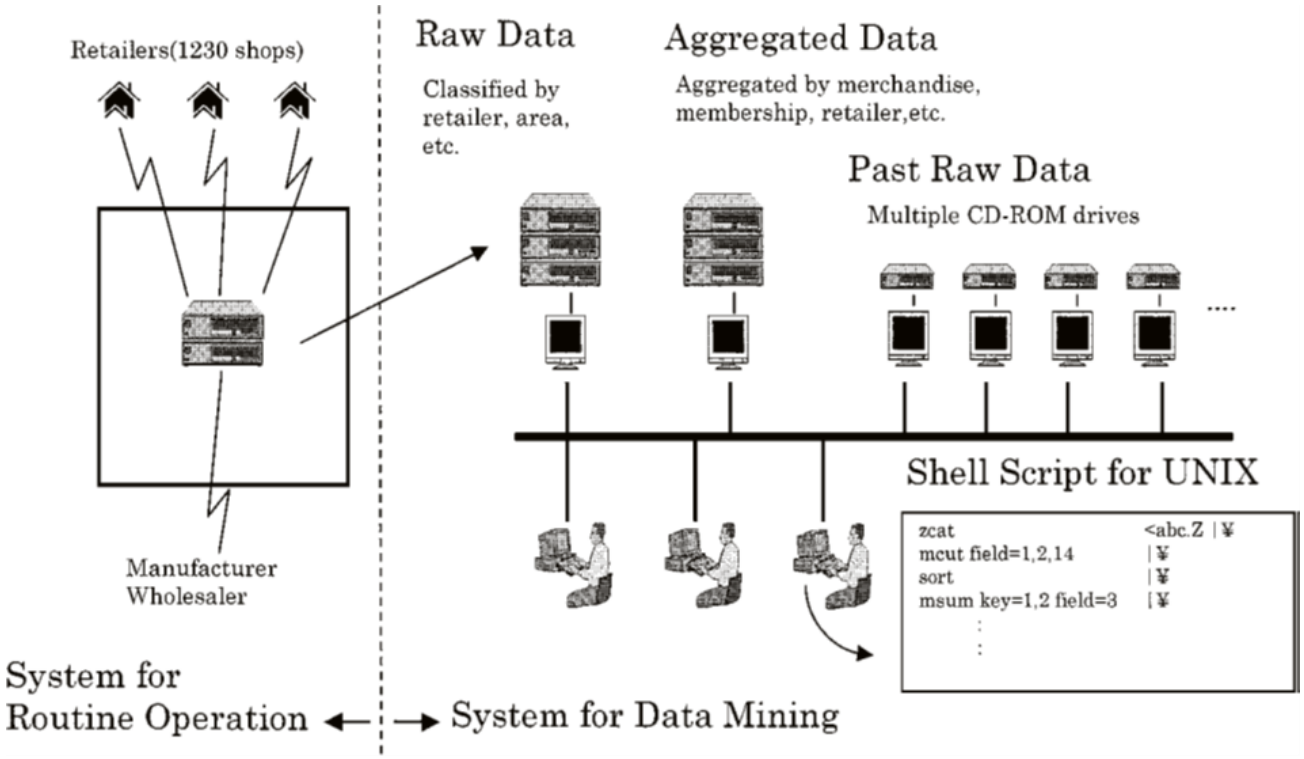

Рисунок 1. *Движение данных о продажах [5]*

Рассмотрим в качестве примера покупку обычного обезболивающего препарата. Обезболивающие средства всегда были первыми кандидатами на продажи со скидкой, поэтому первоначальная мотивация фармацевтической компании была найти способ продавать их по той же цене, но при этом получая большую прибыль. Анализируя данные о продажах аптек с помощью разработанной системой Data Mining было обнаружено, что данное лекарство имеет высокую корреляцию с санитарно-гигиеническими изделиями с точки зрения попутной покупки - покупатели чаще всего брали данную продукцию вместе (заметим, что обнаружить эту корреляцию между обезболивающими и санитарно-гигиеническими изделиями было бы невозможно, если бы входные данные были объединены, например, путем суммирования продаж за единицу от продаж каждого чека). Дальнейший анализ показал, что путем

расположения данных препаратов и изделий рядом, обезболивающие средства могут продаваться в 1,5 раза чаще, чем раньше, даже по обычной цене без скидки.

Генерация ассоциативных правил для фармацевтических данных

Рассмотрим метод ассоциаций, который был применен для выявления корреляции между обезболивающими средствами и санитарно-гигиеническими изделиями. Для поиска правила ассоциаций был использован алгоритм Apriori - алгоритм поиска ассоциативных правил. Данный алгоритм является одним из стандартных алгоритмов нахождения правила ассоциаций среди набора данных [6, 8]. Шаги алгоритма:

1) обнаружение обычных наборов записи;

2) построение ассоциациативных правил на основе найденных наборов.

Рассмотрим часть набора всех медикаментов, продаваемых в аптеках фармацевтической компании (таблица 1) и часть ее аптечных транзакций (таблица 2).

Таблица 1

Описание препаратов (на основе [5])

№	Имя	Группа препарата
1	Витамин D	Витамины
2	Аспирин	Анальгетики
3	Витамин D3	Витамины
4	Сульфанол	Гигиенические средства
5	Амоксициллин	Пенициллины
6	Метронидазол	Противомикробные
7	Пироксикам	Противовоспалительные

Таблица 2

Транзакции (на основе [5])

№	Набор элементов в транзакции
1	{ Витамин D, Аспирин, Витамин D3, Сульфанол }
2	{ Витамин D, Аспирин, Сульфанол }
3	{ Витамин D, Аспирин }
4	{ Аспирин, Витамин D3, Сульфанол }
5	{ Аспирин, Витамин D3}
6	{ Витамин D3, Сульфанол }
7	{ Аспирин, Сульфанол }

На первом шаге алгоритма Apriori необходимо определить минимальную поддержку элемента (в данном примере элементом является препарат) - минимальное число вхождений элемента множества (таблица 1) в транзакции (таблица 2) путем сканирования базы данных. В дальнейшем рассматриваются только элементы и сочетания элементов, которые равны или превышают минимальный порог поддержки (поддержка - число вхождений элемента в транзакцию), так как они считаются часто встречающимися препаратами. Примем минимальный порог поддержки равным трем и подсчитаем ее для элементов таблицы 1 в интеграциях таблицы 2:

Таблица 3

Значение поддержки. Часть 1 (на основе [5])

Набор	Поддержка
{ Витамин D }	3

Набор	Поддержка
{ Аспирин }	6
{ Витамин D3 }	4
{ Амоксициллин }	5
{ Метронидазол }	0
{ Пироксикам }	0

Как мы видим, поддержка элементов {Витамин D}, {Аспирин}, {Витамин D3}, {Сульфанол} превышает минимальный порог, поэтому мы можем создать из них пары и посчитать поддержку для данных пар (таблица 4). Элементы {Метронидазол} и {Пироксикам} не рассматриваются в парах, так как они не прошли минимальный порог поддержки.

Таблица 4

Значение поддержки. Часть 2 (на основе [5])

Набор	Поддержка
{ Витамин D, Аспирин }	3
{ Витамин D, Витамин D3 }	1
{ Витамин D, Амоксициллин }	2
{ Аспирин, Витамин D3 }	3
{ Аспирин, Амоксициллин }	4
{ Витамин D3, Амоксициллин }	3

Пары {Витамин D, Аспирин}, {Аспирин, Витамин D3}, {Аспирин, Амоксициллин}, {Витамин D3, Амоксициллин} превышают минимальный порог 3, поэтому они часто встречаются. Так как {Витамин D, Витамин D3} и {Витамин D, Амоксициллин} не часто встречаются, то любой бóльший набор, который содержит данные наборы тоже не может быть часто встречающимся. Таким образом, мы можем «подрезать» наборы: теперь мы будем искать частые тройки элементов в базе транзакций.

Таблица 5

Значение поддержки. Часть 3

Набор	Поддержка
{ Аспирин, Витамин D3, Сульфанол }	2

Данный набор не является часто встречающимся, а другие наборы были исключены, так как они содержали в себе подмножества с нечасто встречающимися наборами. Таким образом, мы определили все частые наборы элементов в базе данных - {Витамин D}, {Аспирин}, {Витамин D3}, {Амоксициллин}, {Витамин D, Аспирин}, {Аспирин, Витамин D3}, {Аспирин, Амоксициллин}, {Витамин D3, Амоксициллин} - и можем переходить ко второму шагу алгоритма - построению правил ассоциаций.

Ассоциация позволяет выделить устойчивые группы объектов, между которыми существуют неявно заданные связи. Частота появления отдельного предмета или группы предметов, выраженная в процентах, называется распространенностью. Низкий уровень распространенности (менее одной тысячной процента) говорит о том, что такая ассоциация не существенна. Ассоциации записываются в виде правила:

$$A \Longrightarrow B,$$

где: A - посылка, B - следствие. Для определения важности каждого полученного ассоциативного правила необходимо вычислить величину, которую называют поддержкой A к B (или взаимосвязью A и B). Доверие $\sigma(A/B)$ показывает, как часто при появлении A появляется B, и рассчитывается по формуле

$$\sigma(A/B) = \frac{\varepsilon(A \cap B)}{\varepsilon(A)},$$

где: $\varepsilon(A \cap B)$ -поддержка совместного появления A и B; $\varepsilon(A)$ - поддержка A. Например, если $\sigma(A/B) = 20\%$, то это значит, что при покупке товара A в каждом пятом случае приобретается и товар B. Необходимо отметить, что если $\varepsilon(A) \neq \varepsilon(B)$, то $\sigma(A/B) \neq \sigma(B/A)$. В самом деле, покупка компьютера влечет за собой покупку дисков, но покупка дисков не ведет к покупке компьютера. Важной характеристикой ассоциации является мощность, которая рассчитывается по формуле

$$M(A/B) = \frac{\varepsilon(A \cap B)}{\varepsilon(B)}.$$

Чем больше мощность, тем сильнее влияние, которое наличие A оказывает на появление B.

Применим метод ассоциаций к транзакциям из таблицы 2. Для этого рассмотрим набор элементов $A =$ {Аспирин} и $B =$ {Витамин D3}. Так как они не пересекаются, можно обозначить правило ассоциации $R_1 = A \Longrightarrow B$. Поддержка $\varepsilon(A \cap B) = 3$, а поддержка $\varepsilon(A) = 6$. Поэтому доверие данного правила равно

$$R_1 = \frac{\varepsilon(A \cap B)}{\varepsilon(A)} = \frac{3}{6} = 50\%.$$

Из полученного данных следует, что на каждое второе приобретение Аспирин приобретается также {Витамин D3}. Данное правило ассоциации считается значительным, если оно имеет высокую поддержку и высокое доверие. В контексте супермаркета такое правило указывает на то, что клиент, который покупает комплект товаров с большей вероятностью купит данный комплект товаров. Для заданного порога поддержки правила с большими значениями доверия являются более значимыми, чем те, которые имеют меньшее значение доверия. Из определения доверия можно видеть, что оно не больше единицы, поэтому чаще измеряется в процентах, а не в долях единицы.

Анализ групп препаратов в полученных правилах ассоциаций позволили фармацевтической компании найти наиболее часто покупаемые сочетания. Аналитики компании, не имея в своем арсенале систему Data Mining, не смогли бы прийти к такому выводу, так как вручную перебрать такое огромнейшее количество данных просто невозможно.

Заключение

В статье рассмотрена методика использования данных компании для повышения их эффективности и действенности. Показано, как можно использовать алгоритм Apriori для нахождения ассоциативных правил в данных, полученных из аптек фармацевтической компании. Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности, ведь он реагирует на потребности рынка путем разработки мощных информационных технологий. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

ЛИТЕРАТУРА

1. Чубукова И.А., Курс лекций по Data Mining: [Электронный ресурс]. [https://goo.gl/cfvAlQ]. Проверено 19.11.2016.
2. Sakaeda T., Tamon A., Kadoyama K., Okuno Y. Data Mining of the Public Version of the FDA Adverse Event Reporting System. Int J Med Sci 2013; с. 796-803.
3. Sandhya Joshi, Hanumanthachar Joshi. International Journal of Scientific & Engineering Research, Volume 4, Issue 4: Applications of data mining in health and pharmaceutical industry. April-2013. ISSN 2229-5518.
4. Louisville, John Cerrito, Kroger Pharmacy. Survival Data Mining: Treatment of Chronic Illness Patricia Cerrito, SAS Global Forum 2008, Paper 165-2010.
5. Yukinobu H., Naoki K., Yasuyuki M., Katsutoshi Y. Mining Pharmacy Data Helps to Make Profits. DS '98 Proceedings of the First International Conference on Discovery Science, с. 441-442.
6. Discovery of Association Rules in Medical Data. Srinivas Doddi, Achla Marathe (Contact author), S.S. Ravi, David C. Torney. Med Inform Internet Med. 2001 Jan-Mar; с. 25-33.
7. Patricia B. Cerrito, John C. Cerrito, Kroger Pharmacy. Data Mining Methods to Link Multiple Drug Purchases and To Investigate Drug Costs to Consumers. SUGI 29, с. 1-29.
8. Ahmad Y., Fatemeh G.G., Sima A., Somayeh H., Farshad M. Identifying Association Rules among Drugs in Prescription of a Single Drugstore Using Apriori Method. Intelligent Information Management, 2015, с. 253-259.

Pivovarova Natalya Vladimirovna
Bauman Moscow state technical university, Russia, Moscow
pivovarova.natasha2013@yandex.ru

Vidunova Svetlana Igorevna
Bauman Moscow state technical university, Russia, Moscow
E-mail: Svetlana.Vidunova@gmail.com

Data Mining in Pharmaceutical business

Abstract. The growth of data in various areas and the need to analyze them to obtain useful information leads to the fact that many analysts face a variety of challenges. Data collection by itself does not lead to any results, should be considered as raw data to extract useful information. This article discusses the Apriori algorithm for association rules from the set of accumulated data of pharmaceutical companies.

Keywords: data mining; analytical system; Apriori algorithm; association rules; pharmacy; business; profit