УДК 004.056.57

С.А. Нестеров
Санкт-Петербург, Россия

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ СЕРВЕРА УПРАВЛЕНИЯ АНТИВИРУСНОЙ ЗАЩИТОЙ

S.A. Nesterov
St.-Petersburg, Russia

DATA MINING IN A DATABASE OF THE ANTIVIRUS PROTECTION MANAGEMENT SERVER

Рассмотрено использование методов интеллектуального анализа данных для выявления закономерностей в данных, собираемых сервером управления антивирусной защитой. Результаты подобного анализа могут использоваться для более качественного администрирования системы антивирусной защиты.
ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ.

The paper describes some applications of data mining in antivirus protection management systems.
INFORMATION SECURITY. DATA MINING.

Современные корпоративные средства антивирусной защиты, как правило, дополняются средствами централизованного администрирования, которые собирают с клиентских компьютеров большой объем данных. Исследование этих данных методами интеллектуального анализа (data mining) позволяет в ряде случаев выявить в них неявные закономерности, представляющие интерес для администратора безопасности. Эта информация, наряду со стандартными отчетами программного обеспечения (ПО) управления антивирусной защитой, может использоваться для более качественного управления антивирусными клиентами.

В качестве примера рассмотрим такие задачи интеллектуального анализа данных, как кластеризация и классификации.

Кластеризация

Формально задача кластеризации описывается следующим образом [1]. Дано множество объектов данных I, каждый из которых представлен набором атрибутов. Требуется построить множество кластеров C и отображение F множества I на множество C, то есть $F: I \to C$. Количество кластеров может быть заранее неизвестно. Качество решения задачи определяется количеством верно классифицированных (отнесенных к правильному кластеру) объектов данных.

Множество I определим следующим образом:

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\},$$

где i_j – исследуемый объект.

Каждый из объектов характеризуется набором параметров:

$$i_j = \{x_1, x_2, \ldots, x_h, \ldots, x_m\}.$$

Каждая переменная x_n может принимать значения из некоторого множества:

$$x_h = \{v_h^1, v_h^1, \ldots\}.$$

Задача кластеризации состоит в построении множества кластеров:

$$C = \{c_1, c_2, \ldots, c_k, \ldots, c_g\}.$$

где c_k – кластер, содержащий похожие друг на друга (близкие, в соответствии с введенной мерой) объекты из множества I:

$$c_k = \{i_j, i_p \mid i_j \in I, i_p \in I, d(i_j, i_p) < \sigma\},$$

где $d(i_j, i_p)$ – мера близости между объектами, называемая расстоянием. Если расстояние $d(i_j, i_p)$

меньше некоторого значения σ, то считается, что объекты близки, и они помещаются в один кластер.

В рассматриваемой задаче анализа данных, собираемых ПО управления антивирусной защитой, кластеризация может позволить определить группы схожих по параметрам компьютеров, для которых можно ввести единые настройки (или политики) антивирусного клиента.

Допустим, анализ показал наличие группы устаревших компьютеров, пользователи которых работают с программой, создающей файлы большого размера. Проверка подобных файлов резидентным модулем антивирусной защиты существенно увеличивает время открытия и сохранения данных файлов, что вызывает претензии пользователей. Для этой группы компьютеров может быть принято решение об исключении файлов данного типа из списка проверяемых.

Прогнозирование

Формально задачу прогнозирования можно описать следующим образом [1]. Имеется множество объектов:

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\},$$

где i_j – исследуемый объект. Каждый объект характеризуется набором переменных:

$$i_j = \{x_1, x_2, \ldots, x_h, \ldots, x_m, y\},$$

где x_h – независимые переменные (их значения известны), на основании которых надо определить значение зависимой переменной y. Можно обозначить набор независимых переменных вектором X:

$$X = \{x_1, x_2, \ldots, x_h, \ldots, x_m\}.$$

Каждая переменная может принимать значения из некоторого множества. Если множество значений зависимой переменной y – конечное, то речь идет о задаче прогнозирования значения дискретного атрибута, также называемой задачей классификации.

В рассматриваемой предметной области может представлять интерес решение задачи прогнозирования уровня подверженности конкретного компьютера вирусным атакам. Например, при прочих равных условиях, более подвержены заражениям общедоступные компьютеры, которые подключены к Интернет, и на которых разрешено использование съемных носителей. Обучив модель интеллектуального анализа на имеющихся примерах, можно по описанию роли компьютера, его размещения, работающих с ним пользователей, списка установленного ПО и прочих подобных параметров, предсказать, насколько данный компьютер может быть подвержен вирусным инцидентам. В зависимости от результата, можно предпринять какие-то административные меры, например, применить на антивирусном клиенте более или менее жесткую политику.

Экспериментальные результаты

Для того чтобы проверить выдвинутые предположения, применялась база данных (БД) Kaspersky Security Center – средства централизованного администрирования антивирусных продуктов «Лаборатории Касперского». В качестве системы управления базами данных (СУБД) Security Center использует Microsoft SQL Server, поэтому было решено использовать встроенные средства интеллектуального анализа данных этой СУБД. В работе использовался MS SQL Server 2008 R2 в редакции developer.

Решение задачи кластеризации в MS SQL Server 2008 может выполняться с помощью алгоритмов *k*-средних («жесткая» кластеризация, каждый элемент относится только к одному кластеру) и максимизации ожидания («мягкая» кластеризация, элемент может относиться к нескольким кластерам с разными вероятностями) [2]. Задача классификации может быть решена алгоритмами деревьев решений, нейронных сетей, логистической регрессии (в реализации Microsoft это вариант алгоритма нейронных сетей, формирующий сеть без скрытого слоя), упрощенным алгоритмом Байеса. Также данная задача может быть решена с помощью алгоритма кластеризации, когда значение прогнозируемого атрибута сначала рассматривается как неопределенное, а после выбирается наиболее характерное значение для кластера, к которому отнесен данный элемент.

В ходе выполнения работы был проведен анализ данных, собираемых сервером управления антивирусной защитой Kaspersky Security Center 9.0, развернутым в сети кафедры системного анализа и управления Санкт-Петербургского государственного политехнического университета. Защищаемая сеть включает два компьютерных класса, компьютеры преподавателей, три постоянно работающих сервера и два сервера, включае-

мых на время лабораторных работ. Расположение компьютеров частично отображено на структуру групп администрирования Security Center.

На управляемых Security Center компьютерах сети используется антивирусное ПО Kaspersky Endpoint Security 8.0. Кроме этого, в сети присутствует около 20 кафедральных компьютеров с антивирусным ПО других производителей, а также периодически подключаемые в кафедральную сеть личные ноутбуки сотрудников и студентов. Информация о них также попадает в базу, но является неполной, и это нужно учитывать в процессе анализа (записи о таких хостах отфильтровывались при обучении и анализе модели). Анализируемая БД содержит записи о 118 хостах, 29 из которых управляются Security Center.

Нужно отметить, что полная документация на БД Security Center пользователям не предоставляется. Для конечного пользователя предназначены формируемые этим ПО стандартные отчеты, которые достаточны для обычных задач администрирования, но в рассматриваемом случае малопригодны. Но в этой БД определены несколько «публичных» представлений (view), документация на которые поставляется с продуктом. Данные этих представлений и использовались при определении источника данных для анализа. К ним относятся:

данные о группах компьютеров (группах администрирования Security Center) и их иерархии;

сведения о компьютерах (хостах) – имя, домен, операционная система, дата и время, когда хост был доступен, дата и время последнего обновления антивируса, и другие подобные данные;

информация о текущем статусе хоста с точки зрения антивирусной защиты – идентификатор статуса (0 – OK; 1 – Warning; 2 – Critical) и расширенное описание состояния антивирусного ПО и агента администрирования антивирусного ПО;

перечень событий на хостах – обнаружен вирус, обнаружена сетевая атака, инфицированный объект вылечен и т. д.;

информация о приложениях, установленных на управляемых хостах, их обновлениях, используемой аппаратуре.

При выполнении интеллектуального анализа данных с помощью аналитических служб MS SQL Server можно сначала создать структуру интеллектуального анализа, а потом в ее рамках – одну или несколько моделей. В структуре описывается перечень анализируемых параметров, их тип данных и тип содержимого (например, числовое непрерывное или числовое дискретное значение). Определение модели включает указание на используемый алгоритм, его параметры, фильтры, применяемые к данным этой модели. После этих определений исходные данные загружаются в структуру и появляется возможность обработать их с помощью алгоритма интеллектуального анализа – обучить модель. Выявленные в результате обучения закономерности сохраняются в модели и могут использоваться для последующего анализа и прогнозирования. При необходимости часть данных может резервироваться в структуре для целей тестирования – они не используются при обучении модели.

Для анализа исходного набора с помощью алгоритма кластеризации создана структура данных и в ней определено несколько моделей, одни из которых использовали алгоритм максимизации ожидания, другие – *k*-средних, с разными параметрами. В настройках алгоритмов устанавливалось число формируемых кластеров равное 2, 3, 4, 5, 6 и 0 (число кластеров определяется алгоритмом эвристически). Явно задавать более шести кластеров не представлялось целесообразным из-за малого объема исходных данных: число управляемых Security Center компьютеров в исследованной БД равно 29.

По результатам сделан выбор в пользу алгоритма максимизации ожидания с формированием четырех кластеров. Визуальный анализ результатов обучения модели можно сделать с помощью ряда встроенных графических средств. В частности, это диаграммы «профилей» кластеров, показывающие характерные для кластера значения каждого из атрибутов. Один из таких атрибутов – число обнаруженных вирусных инцидентов nVirusCount (данные брались за период времени около трех месяцев). Фрагмент диаграммы профилей кластеров для выбранной модели приведен на рис. 1.

На рисунке видно, что в кластерах 2 и 4 оказались компьютеры с низким (менее трех) числом вирусных инцидентов, а в кластерах 1 и 3 – число заражений обычно больше. При этом невылеченные объекты (nUncured) имеются только на компьютерах, принадлежащих кластеру 1.

В связи с тем что задача кластеризации является описательной, большинство средств оценки точности к моделям кластеризации неприменимы. Из предоставляемых MS SQL Server 2008

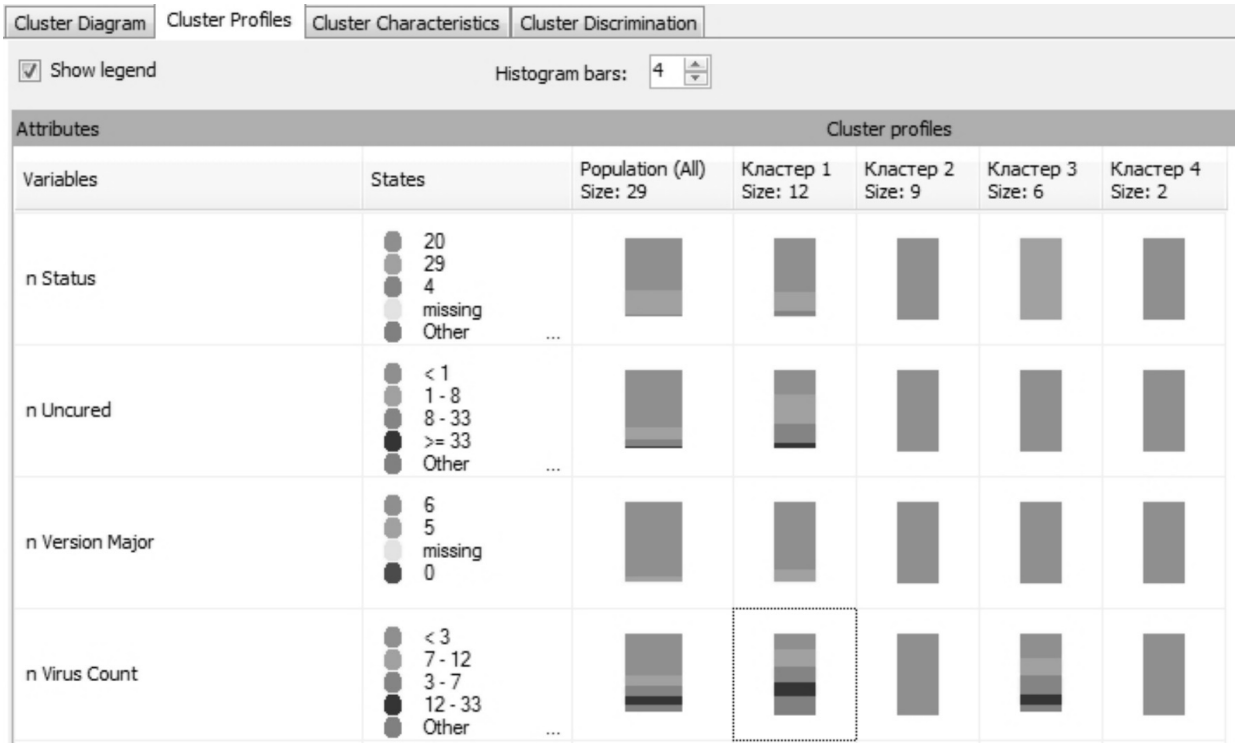


Рис. 1. Диаграмма профилей кластеров

инструментов можно воспользоваться только перекрестной проверкой (cross validation). Суть ее заключается в том, что множество вариантов, которое использует модель, разбивается на непересекающиеся подмножества (разделы). Для каждого из разделов производится обработка модели и полученные результаты сравниваются с теми, что были на исходном множестве вариантов. Если результаты близки, можно говорить об удачной модели интеллектуального анализа: результат прогноза или анализа достаточно стабилен. На рис. 2 представлены результаты перекрестной проверки для случая, когда исходное множество вариантов разбивалось на три подмножества.

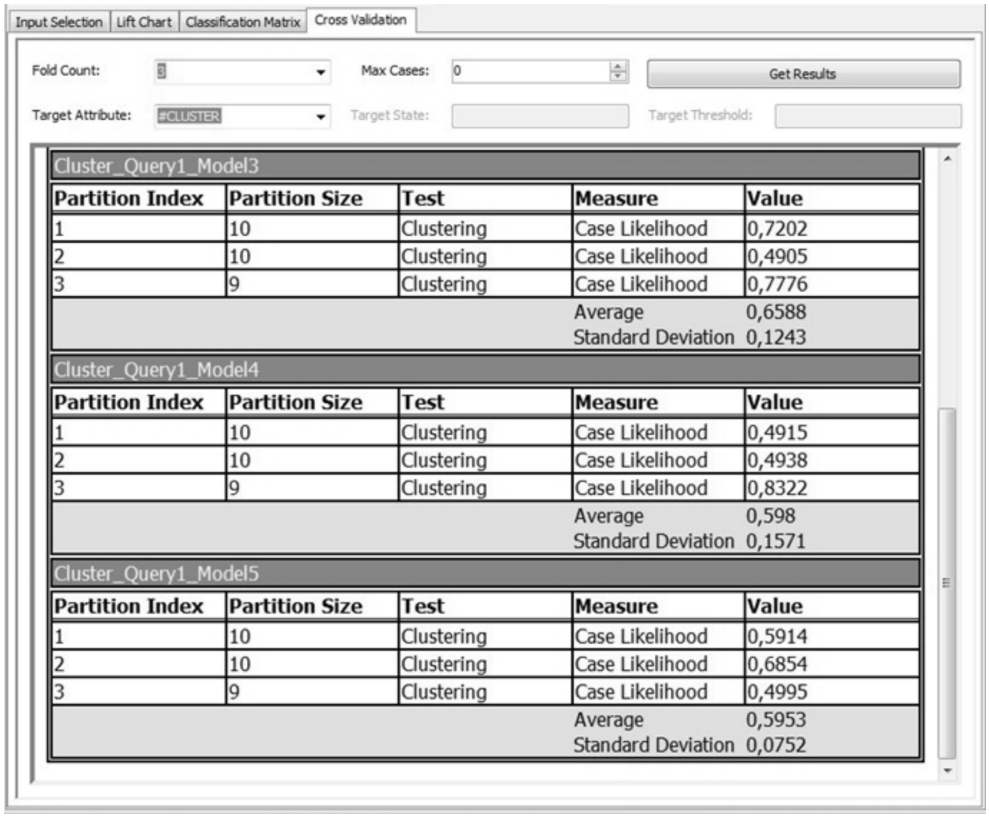
Использование большего числа разделов не представляется обоснованным из-за малого объема исходных данных. Из рисунка видно, что выбранная модель Cluster_Query1_Model3 показывает наилучший результат – около 66 % совпадений в результатах кластеризации исходного множества и в рамках разделов. Здесь также следует учитывать, что для анализа имелось всего 29 элементов.

Кроме редакторов интегрированной среды разработки BI Development Studio, работа с моделью может производиться и с помощью SQL-подобного языка DMX (data mining eXtensions).

Запрос, позволяющий получить информацию о компьютерах, которые были отнесены к первому кластеру (условие IsInNode('001')), будет выглядеть следующим образом (выводится имя, число вирусных инцидентов, название группы администрирования и подпись «Cluster1»):

```
SELECT StructureColumn('Wstr Win Name') as HostName, [n Virus Count], [Wstr Name] as GrName, 'Cluster1' as Clust
FROM Cluster_Query1_Model3.CASES
WHERE IsInNode('001')
```

Анализ полученного списка хостов показал, что в этот кластер (наименее благополучный по числу вирусных инцидентов) попала большая часть компьютеров из учебного класса, где в основном занимаются студенты младших курсов. Результат можно объяснить следующим образом: за каждым компьютером регулярно работает много пользователей, часто используются съемные носители. В данном классе можно рекомендовать установить более жесткую политику для антивирусного клиента (проверка в режиме реального времени архивов, более частое полное сканирование системы, ограничение работы со съемными накопителями).



Cluster_Query1_Model3

Partition Index	Partition Size	Test	Measure	Value
1	10	Clustering	Case Likelihood	0,7202
2	10	Clustering	Case Likelihood	0,4905
3	9	Clustering	Case Likelihood	0,7776
			Average	0,6588
			Standard Deviation	0,1243

Cluster_Query1_Model4

Partition Index	Partition Size	Test	Measure	Value
1	10	Clustering	Case Likelihood	0,4915
2	10	Clustering	Case Likelihood	0,4938
3	9	Clustering	Case Likelihood	0,8322
			Average	0,598
			Standard Deviation	0,1571

Cluster_Query1_Model5

Partition Index	Partition Size	Test	Measure	Value
1	10	Clustering	Case Likelihood	0,5914
2	10	Clustering	Case Likelihood	0,6854
3	9	Clustering	Case Likelihood	0,4995
			Average	0,5953
			Standard Deviation	0,0752

Рис. 2. Результаты перекрестной проверки

В приведенном выше примере найденную зависимость можно было выявить и без использования data mining: администратор мог сделать соответствующее предположение и проверить его, просмотрев отчеты. Но здесь важно отметить, что перед проведением анализа средствами data mining никаких предварительных версий не было, и здесь работа алгоритмов анализа отчасти заменила профессиональную интуицию специалиста. Таким образом, даже для небольшого объема данных проведение кластеризации и анализ ее результатов позволили выявить полезные зависимости.

Аналогичный эксперимент с прогнозированием дискретного атрибута получился менее удачным. В процессе обучения основанной на алгоритме деревьев решений модели было получено сообщение, что «алгоритм не обнаружил разбиений для прогнозирующей модели». Другие алгоритмы (нейронных сетей, логистической регрессии, упрощенный алгоритм Байеса) по итогам тестирования показали низкую точность предсказания. По всей видимости, в данном случае сказался небольшой объем обучающей выборки: всего в используемом наборе находились данные о 29 компьютерах, причем часть из них резервировалась для целей тестирования.

Результаты работы показывают, что методы интеллектуального анализа данных позволяют выявить неявные зависимости, полезные для администратора антивирусной защиты. В частности, решение задачи кластеризации может указать на наиболее адекватный способ распределения компьютеров по группам администрирования.

Использование прогнозирующих моделей позволит, в подобном случае, отнести новый компьютер в наиболее подходящую группу администрирования. Но для обучения прогнозирующей модели необходима достаточная обучающая выборка данных, которую можно получить только в сети, включающей большое число защищаемых компьютеров. Анализ решения сходных задач, описанных в литературе [2, 3], позволяет надеяться на получение в подобном случае положительных результатов.

Работа проведена при поддержке ЗАО «Лаборатория Касперского» в форме гранта в рамках Программы поддержки инновационных проектов.

СПИСОК ЛИТЕРАТУРЫ

1. **Барсегян, А.А.** Анализ данных и процессов: Учеб. пособие [Текст] / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. –3-е изд., перераб. и доп. –СПб.: БХВ-Петербург, 2009. –512 с.

2. **Макленнен, Джеми.** Microsoft SQL Server 2008: Data mining – интеллектуальный анализ данных [Текст] / Джеми Макленнен, Чжаохуэй Танг, Богдан Криват; Пер. с англ. –СПб.: БХВ-Петербург, 2009. –720 с.

3. **Аникин, И.В.** Технология интеллектуального анализа данных для выявления внутренних нарушителей в компьютерных системах [Текст] / И.В. Аникин // Научно-технические ведомости СПбГПУ. –2010. –№ 6 (113). –С. 112–117.