Data Mining - Chuong 2 - Tài liệu môn khai thác dữ liệu

PHÂN TÍCH DỮ LIỆU spss (Trường Đại học Kinh tế Thành phố Hồ Chí Minh)

CHƯƠNG 2

QUY TRÌNH KHAI THÁC DỮ LIỆU VÀ KHÁM PHÁ TRI THỨC

Nội dung chương này:

- Mô tả Quy trình chuẩn liên ngành của Khai thác dữ liệu (Cross-industry Standard for Data mining: CRISP-DM) - một tập hợp các giai đoạn (phrase) trong nghiên cứu khai thác dữ liệu.
- Thảo luận chi tiết từng giai đoạn.
- Một số ví dụ minh họa.
- Thảo luận về quy trình khám phá tri thức.

Khi phải đối mặt với một lượng lớn dữ liệu, doanh nghiệp (hoặc các tổ chức) hưởng lợi nhờ các quy trình có tính hệ thống làm cho những bộ dữ liệu trở nên ý nghĩa. Ví dụ, một doanh nghiệp cung cấp các dịch vụ cho khách hàng và gặp phải hiện tượng thanh toán hóa đơn không đúng kỳ hạn. Có những đánh đổi khi xử lý những món nợ chậm trả này. Nếu cắt dịch vụ ngay lần đầu tiên thanh toán trễ sẽ là động cơ cho khách hàng trả tiền đúng hạn. Tuy nhiên, cũng là động cơ ngăn cản việc sử dụng dịch vụ. Nếu công ty cung cấp dịch vụ điện thoại ngắt kết nối ngay ngày hôm sau khi hóa đơn chưa được thanh toán, công ty này sẽ mất rất nhiều khách hàng, và điều này sẽ ảnh hưởng tiêu cực tới lợi nhuận của công ty. Mặt khác, không cần thiết phải phung phí tiền bạc để theo dõi những tài khoản chưa thanh toán trong vòng 5 năm. Thật không dễ để trả lời câu hỏi hóa đơn có thể chậm bao lâu trước khi khách hàng bị xóa tên. Trên thực tế, một số loại khách hàng được đối xử khác với những loại khách hàng khác. Khai thác dữ liệu cung cấp công cụ để giúp nhận ra mức độ ảnh hưởng của các chính sách khác nhau, vì vậy cho phép công ty thực hiện những chính sách hợp lý hơn.

Để thực hiện các phân tích khai thác dữ liệu cần có một quy trình tổng quát. Chương này mô tả một quy trình chuẩn thường được sử dụng, gồm một chuỗi các bước thường thấy trong nghiên cứu khai thác dữ liệu. Tuy không nhất thiết phải có tất cả các bước trong mỗi phân tích, nhưng quy trình này cung cấp một mức độ bao phủ tốt các bước cần thiết, bắt đầu bằng khảo sát dữ liệu, thu thập dữ liệu, xử lý dữ liệu, phân tích, rút ra kết luận, và triển khai.

CRISP-DM

Quy trình chuẩn liên ngành cho Khai thác dữ liệu (Cross-industry Standard for Data mining: CRISP-DM) được sử dụng rộng rãi. Mô hình này gồm 6 giai đoạn của một quy trình có tính chu kỳ.

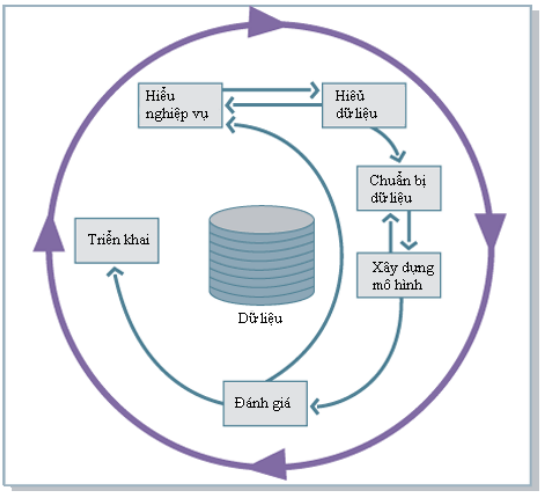
- **Hiểu yêu cầu (Business understanding):** hiểu yêu cầu bao gồm việc xác định mục tiêu kinh doanh, đánh giá hiện trạng, thiết lập mục tiêu khai thác dữ liệu, và xây dựng kế hoạch dự án.
- **Hiểu dữ liệu (Data understanding):** một khi mục tiêu kinh doanh và kế hoạch dự án được thiết lập, hiểu dữ liệu liên quan đến yêu cầu của dữ liệu. Bước này bao gồm thu thập dữ liệu ban đầu, mô tả dữ liệu, khảo sát dữ liệu, và kiểm tra chất lượng dữ liệu. Khảo sát dữ liệu như là xem những tóm tắt thống kê (bao gồm biểu diễn trực quan của biến phân loại) có thể xuất hiện vào cuối giai đoạn này. Những mô hình như phân tích cụm cũng có thể áp dụng trong giai đoạn này với chủ ý nhận dạng mô hình (patterns) dữ liệu.
- **Chuẩn bị dữ liệu (Data preparation):** khi nguồn dữ liệu được xác định, dữ liệu cần được chọn lọc, làm sạch, gắn vào các mẫu (form) mong muốn, và định dạng. Làm sạch và biến

đổi dữ liệu theo mô hình yêu cầu xảy ra trong giai đoạn này. Khảo sát dữ liệu ở mức độ sâu hơn có thể áp dụng trong giai đoạn này, các mô hình bổ sung có thể được sử dụng, tăng khả năng tìm ra mô hình dựa vào việc hiểu yêu cầu.

- **Xây dựng mô hình (Modeling):** các công cụ phần mềm khai thác dữ liệu như trực quan hóa (visualization) (vẽ sơ đồ dữ liệu và thiết lập các mối quan hệ) và phân tích nhóm (cluster analysis) (để xác định các biến có thể đi cùng với nhau) rất hữu ích cho phân tích sơ khởi. Các công cụ như quy nạp tổng quát có thể xây dựng các quy tắc kết hợp ban đầu. Một khi hiểu hơn về dữ liệu (thường thông qua nhận dạng mô hình khởi đầu bằng quan sát các mô hình đầu ra) nhiều mô hình chi tiết hợp lý cho dữ liệu có thể được ứng dụng. Phân chia dữ liệu thành tập phân tích (training set) và tập kiểm tra (test set) cũng cần thiết cho xây dựng mô hình.
- **Đánh giá mô hình (Evaluation):** các kết quả mô hình cần được đánh giá trong bối cảnh mục tiêu kinh doanh được thiết lập ở giai đoạn đầu tiên (hiểu yêu cầu). Điều này sẽ đưa đến việc xác định những nhu cầu khác (thường thông qua các mô hình nhận dạng), nên thường xuyên quay lại những giai đoạn trước đó của quy trình CRISP-DM. Hiểu được yêu cầu kinh doanh là một thủ tục lặp trong khai thác dữ liệu, nơi mà các kết quả khác nhau trong các công cụ trực quan hóa, thống kê, và trí tuệ nhân tạo chỉ ra cho người phân tích các mối quan hệ mới sẽ giúp hiểu biết sâu sắc hơn về hoạt động của tổ chức.
- **Triển khai (Deployment):** khai thác dữ liệu có thể được sử dụng cho cả việc kiểm định giả thuyết đã thiết thiết lập, và cả khám phá tri thức (nhận ra những mối quan hệ hữu ích bất ngờ). Thông qua việc tri thức được khám phá trong các giai đoạn trước đó của quy trình CRISP-DM, các mô hình đạt được có thể áp dụng cho hoạt động kinh doanh trong nhiều mục đích, bao gồm dự báo hoặc nhận dạng các tình hình quan trọng. Những mô hình này cần được giám sát với các thay đổi trong điều kiện vận hành doanh nghiệp, bởi vì nó có thể đúng hôm nay nhưng không còn đúng trong năm tới. Nếu một thay đổi lớn xảy ra, có thể phải xây dựng lại mô hình. Ghi nhận lại các kết quả của dự án khai thác dữ liệu và tài liệu hóa các chứng cứ cho các nghiên cứu trong tương lai là điều khôn ngoan.

Hình 2.1 phác họa quy trình này. Quy trình 6 bước này không phải là cứng nhắc bởi các số thứ tự của quy trình. Có nhiều giai đoạn được quay lại. Thêm vào đó, những chuyên gia có kinh nghiệm có thể không cần áp dụng tất cả các giai đoạn cho các nghiên cứu. Nhưng CRISP-DM cung cấp một khung quy trình hữu ích cho khai thác dữ liệu.

Hình 2.1 Quy trình CRISP-DM cho khai thác dữ liệu.



Hiểu yêu cầu ứng dụng trong kinh doanh

Yếu tố quan trọng của nghiên cứu khai thác dữ liệu là phải biết được nghiên cứu để làm gì. Điều này bắt đầu từ nhu cầu quản lý đòi hỏi tri thức mới, và diễn đạt mục tiêu kinh doanh

thành công việc nghiên cứu. Mục đích của các câu hỏi như "Loại khách hàng nào quan tâm đến mỗi loại sản phẩm của chúng ta?" hay "Hồ sơ của các khách hàng điển hình, và giá trị của họ đối với chúng ta như thế nào?" là cần thiết. Sau đó một kế hoạch tìm kiếm những nhu cầu kiến thức này được xây dựng về các mặt thu thập dữ liệu, phân tích dữ liệu, và báo cáo. Trong giai đoạn này, một ngân sách hỗ trợ việc nghiên cứu cần được thiết lập, chí ít trong thời kỳ bắt đầu.

Trong các mô hình phân khúc khách hàng, như kinh doanh bán lẻ bằng catalog của Fingerhut, nhận dạng mục đích kinh doanh có ý nghĩa là nhận dạng loại khách hàng nào được kỳ vọng giúp nâng cao suất sinh lợi. Phân tích cho các nhà phân phối thẻ tín dụng cũng hữu ích tương tự. Với mục tiêu kinh doanh, cửa hàng bách hóa thường cố gắng xác định món hàng nào đươc mua chung với nhau để sắp xếp theo "bố trí tương thích" (affinity positioning) trong cửa hàng, hay để sử dụng thông minh cho những chiến dịch khuyến mãi. Khai thác dữ liệu có nhiều ứng dụng hữu ích trong kinh doanh, một số ứng dụng sẽ được giới thiệu qua các bài học trong suốt cuốn sách này.

Hiểu dữ liệu

Vì khai thác dữ liệu có tính định hướng tác vụ (task-oriented), những tác vụ kinh doanh khác nhau đòi hỏi những bộ dữ liệu khác nhau. Giai đoạn đầu tiên của quy trình khai thác dữ liệu là lựa chọn dữ liệu liên quan từ nhiều cơ sở dữ liệu để mô tả một cách đúng đắn nhiệm vụ kinh doanh. Có ít nhất 3 vấn đề cần lưu ý khi chọn dữ liệu. Vấn đề thứ nhất là thiết lập bản mô tả vấn đề chính xác và rõ ràng. Ví dụ, dự án khai thác dữ liệu trong kinh doanh bán lẻ nhằm xác định hành vi tiêu dùng của khách hàng nữ trong việc mua sắm áo quần theo mùa. Một ví dụ khác là xác định mô hình vỡ nợ của những người sử dụng thẻ tín dụng. Vấn đề thứ hai là xác định dữ liệu liên quan đến vấn đề được mô tả. Hầu hết dữ liệu nhân khẩu, giao dịch thẻ tín dụng, và tài chính có thể liên quan đến cả dự án khai thác dữ liệu ngành bán lẻ và vỡ nợ thẻ tín dụng. Tuy nhiên, dữ liệu về giới tính có thể bị luật pháp cấm sử dụng mới đây, nhưng lại là hợp pháp và quan trọng trong thời gian trước. Vấn đề thứ ba là các biến được chọn cho những dữ liệu liên quan phải độc lập với nhau. Biến độc lập có nghĩa là các biến này không chứa những thông tin trùng lập nhau. Sự chọn lọc cẩn thận các biến độc lập có thể làm cho các thuật toán khai thác dữ liệu nhanh chóng tìm ra các mô hình.

Nguồn dữ liệu cho việc chọn lọc dữ liệu rất đa dạng. Thông thường, các loại nguồn dữ liệu cho ứng dụng trong kinh doanh bao gồm **dữ liệu nhân khẩu** (như thu nhập, giáo dục, số người trong hộ gia đình, và tuổi), **dữ liệu xã hội** (như sở thích, thành viên câu lạc bộ, và giải trí), **dữ liệu giao dịch** (ghi nhận mua bán, chi tiêu thẻ tín dụng, séc đã phát hành, vv…). Dữ liệu định lượng đo đạc được bởi các con số. Nó có thể hoặc rời rạc (như số nguyên) hoặc liên tục (như số thực). **Dữ liệu định tính**, còn được biết đến là dữ liệu phân loại, gồm hai loại dữ liệu danh nghĩa (nominal) và dữ liệu thứ tự (ordinal). Dữ liệu danh nghĩa giới hạn không có giá trị thứ tự, như là dữ liệu giới tính chỉ có hai giá trị: nam hoặc nữ. Dữ liệu thứ tự có giá trị theo thứ tự. Ví dụ, xếp hạng tín dụng của khách hàng được xem như dữ liệu thứ tự bởi vì xếp hạng có thể là xuất sắc, khá và xấu. Dữ liệu định lượng có thể dễ dàng biểu diễn bằng các loại phân phối xác suất. Phân phối xác suất mô tả dữ liệu phân tán hay tập trung như thế nào. Ví dụ, dữ liệu thông thường phân phối đối xứng và thường có dạng hình chuông. Dữ liệu định tính có thể bị mã hóa thành số và sau đó mô tả bằng tần suất xác suất. Một khi dữ liệu liên quan được chọn lựa dựa vào khai thác dữ liệu theo mục đích kinh doanh, chuẩn bị dữ liệu cần được thực hiện.

Chuẩn bị dữ liệu

Mục đích của tiền xử lý dữ liệu là làm sạch những dữ liệu đã chọn lọc để có được chất lượng dữ liệu tốt hơn. Các dữ liệu được chọn lọc có thể khác nhau về định dạng bởi vì chúng được chọn từ những nguồn dữ liệu khác nhau. Nếu dữ liệu có dạng tập tin phẳng, tin thoại, văn bản web, chúng cần phải được chuyển đổi sang định dạng điện tử thống nhất. Một cách tổng quát,

làm sạch dữ liệu có nghĩa là lọc, gộp, và điền giá trị mới cho các giá trị khuyết (imputation). Bằng cách lọc dữ liệu, những dữ liệu chọn lọc được kiểm tra để tìm ra các dữ liệu bất thường (outlier) và dữ liệu thừa. Dữ liệu bất thường khác biệt rất lớn với đa số dữ liệu khác, hoặc là dữ liệu hoàn toàn nằm ngoài nhóm dữ liệu đã chọn lọc. Ví dụ, nếu thu nhập của một khách hàng trung lưu là 250.000 USD, đây là một lỗi và cần phải loại ra khỏi dự án khai thác dữ liệu – dự án khảo sát những khía cạnh khác nhau của tầng lớp trung lưu. Dữ liệu bất thường có thể được tạo ra bởi nhiều nguyên nhân, có thể do lỗi của con người hoặc do lỗi kĩ thuật, hoặc có thể xảy ra một cách tự nhiên do các sự kiện đặc biệt. Giả sử tuổi của chủ thẻ tín dụng được ghi nhận là "12". Điều này do lỗi của con người. Tuy nhiên, thực tế có thể có những thiếu niên sống độc lập, giàu có và thích mua sắm. Xóa bỏ các dữ liệu ngoại biên một cách tùy tiện có thể làm mất những thông tin có giá trị.

Dữ liệu thừa là những dữ liệu được ghi nhận trùng lặp theo nhiều cách. Doanh số bán hàng hàng ngày của mỗi sản phẩm là dư thừa đối với doanh số theo mùa của chính sản phẩm đó, bởi vì chúng ta có thể tìm được doanh số của công ty từ doanh số mỗi ngày hoặc từ doanh số theo mùa. Đối với dữ liệu gộp, kích thước dữ liệu có thể được giảm nhờ vào các thông tin gộp. Lưu ý rằng cho dù tập dữ liệu gộp có thể nhỏ, thông tin đó vẫn phải được giữ lại. Nếu một chương trình tiếp thị cho việc bán hàng nội thất được xem xét trong 3 hoặc 4 năm tới, dữ liệu bán hàng mỗi ngày có thể được gộp thành dữ liệu năm. Kích thước của dữ liệu bán hàng sẽ giảm xuống một cách đáng kể. Bằng việc làm nhẵn dữ liệu, những giá trị bị khuyết của dữ liệu được chọn được tìm thấy và những giá trị mới hoặc giá trị thích hợp sẽ được bổ sung thay chỗ khuyết. Những giá trị này có thể là giá trị trung bình của các biến hoặc mode. Một giá trị khuyết có thể không đem lại kết quả khi thuật toán khai thác dữ liệu được ứng dụng để tìm kiếm các mô hình tri thức.

Dữ liệu có thể được biểu diễn thành nhiều dạng khác nhau. Ví dụ, trong CLEMENTINE, những dạng dữ liệu sau có thể được sử dụng:

- Range (dạng dãy): trị số (số nguyên, số thực, hoặc ngày/giờ).
- Flag (dạng cờ): nhị phân - có/không, 0/1, hoặc dữ liệu có 2 biểu hiện (văn bản, số nguyên, số thực, hoặc ngày/giờ).
- Set (dạng bộ) : dữ liệu với nhiều giá trị xác định (dữ liệu số, chuỗi, hoặc ngày/giờ).
- Typeless (không có dạng): cho những dạng dữ liệu khác.

Thông thường chúng ta nghĩ rằng dữ liệu là số thực, như tuổi tác, thu nhập hàng năm tính bằng đô-la (chúng ta sẽ sử dụng dạng dữ liệu dãy cho các trường hợp này). Đôi khi các biến xuất hiện nhiều dạng, ví dụ như có bằng lái xe hay không, hoặc bồi thường bảo hiểm có gian lận hay không. Thường hợp này có thể xử lý bằng việc sử dụng giá trị thực (như 1 hoặc 0). Nhưng sẽ hữu hiệu hơn nếu sử dụng biến Flag. Thông thường, sẽ phù hợp hơn khi xử lý với dữ liệu phân loại, như là tuổi dưới dạng tập hợp của [trẻ, trung niên, già], hoặc thu nhập dưới dạng tập hợp của [thấp, trung bình, cao]. Trong trường hợp đó, chúng ta có thể nhóm dữ liệu lại và gán một phân loại thích hợp dữ liệu dưới dạng chuỗi thành một dữ liệu tập hợp. Dạng toàn diện nhất là "dãy" (range), nhưng đôi khi dữ liệu không ở dạng đó nên các nhà phân tích buộc phải sử dụng dạng tập hợp (set) hoặc (flag). Đôi khi có thể chính xác hơn nếu xử lý với dữ liệu dạng tập hợp (set) hơn là dữ liệu dạng dãy (range).

Một ví dụ khác, chương trình PolyAnalyst cho các loại dữ liệu sau:

- Số (numerical): trị số liên tục
- Nguyên (integer): trị số nguyên
- Có/không (yes/no): dữ liệu nhị phân
- Phân loại (category): một tập hợp các trị số có thể
- Ngày (date)

- Chuỗi (string)
- Văn bản (text)

Mỗi phần mềm có một kiểu dữ liệu khác nhau, nhưng những dạng dữ liệu chính được giới thiệu trong 2 danh sách trên.

Có nhiều phương pháp thống kê và công cụ trực quan hóa có thể được sử dụng để tiền xử lý dữ liệu được chọn. Thống kê phổ biến, như là giá trị cực đại, cực tiểu, trung bình, mode có thể có thể sử dụng để gộp hoặc làm nhẵn dữ liệu, biểu đồ phân tán và biểu đồ hình hộp thường được sử dụng để lọc các dữ liệu bất thường. Nhiều kĩ thuật cao hơn (như phân tích hồi quy, phân tích cụm, cây quyết định, hay phân tích phân cấp) có thể được ứng dụng trong tiền xử lý dữ liệu dựa vào yêu cầu về chất lượng của dữ liệu được chọn. Bởi vì tiền xử lý dữ liệu khá tỉ mỉ và dông dài, nó đòi hỏi rất nhiều thời gian. Trong một số trường hợp, tiền xử lý dữ liệu có thể tiêu tốn 50% tổng thời gian của quy trình khai thác dữ liệu. Dữ liệu đơn giản và có định dạng chuẩn có được từ giai đoạn tiền xử lý có thể chia sẻ thông qua nhiều hệ thống máy tính khác nhau, có thể tạo ra sự linh hoạt trong triển khai những thuật toán khai thác dữ liệu hoặc công cụ khai thác dữ liệu.

Là một thành phần quan trọng trong khâu chuẩn bị dữ liệu, biến đổi dữ liệu là sử dụng những công thức toán học đơn giản hoặc đường cong học hỏi (learning curves) để chuyển đổi những phép đo lường khác nhau của dữ liệu được chọn đã làm sạch về một thang đo thống nhất cho mục đích phân tích dữ liệu. Có rất nhiều phép đo lường thống kê, như trung bình, trung vị, mode và phương sai có thể được sử dụng để chuyển đổi dữ liệu. Về mặt biểu diễn dữ liệu, biến đổi dữ liệu có thể dùng để (1) biến đổi dữ liệu số sang thang đo số (numerical scale), và (2) mã hóa lại dữ liệu phân loại thành thang đo số. Trường hợp dữ liệu số thành thang đo số, chúng ta có thể sử dụng các biến đổi toán học để làm "co lại" hoặc "mở rộng" dữ liệu được cho. Một lý do để biến đổi là để khử những khác biệt trong thang đo biến. Lấy ví dụ, thuộc tính "lương" có giá trị trong khoảng 20.000$ đến 70.000$, chúng ta có thể sử dụng công thức S = (x-min)/(max-min) để làm "co lại" các giá trị lương đã biết, như 50.000$ trở thành 0.6, thành một số trong khoảng [0;1]. Nếu giá trị trung bình của lương là 45.000$, và độ lệch chuẩn được cho là 15.000$, mức lương 50.000$ có thể biến đổi thành 0.33. Biến đổi dữ liệu từ hệ mét (metric system) (ví dụ như mét, kilomet) sang hệ đo lường của Anh (như bộ, dặm) cũng là một ví dụ. Trường hợp dữ liệu phân loại thành thang đo số, chúng ta phải gán một giá trị số thích hợp cho giá trị phân loại phụ thuộc vào nhu cầu. Biến phân loại có thể có tính thứ tự (như: yếu, trung bình, và mạnh) và có tính danh nghĩa (như: đỏ, vàng, xanh dương, và xanh lục). Ví dụ, một biến nhị phân {có, không} có thể biến đổi thành "1=có và 0=không". Lưu ý rằng biến đổi một giá trị số thành một giá trị thứ bậc có nghĩa là biến đổi với thứ tự, trong khi biến đổi thành một giá trị danh nghĩa là một biến đổi ít cứng nhắc hơn. Chúng ta cần phải cẩn thận không sử dụng các biến đổi tạo ra cấp dữ liệu cao hơn dữ liệu gốc. Ví dụ, thang đo Likert thường trình bày các thông tin có tính thứ tự được mã hóa thành số (1 đến7, 1 đến 5, và vân vân). Tuy nhiên, những con số này không ngụ ý một thanh đo thông thường về sự sai khác. Một thứ được đánh giá là 4 không hẳn có nghĩa là có ý nghĩa gấp đôi so với thứ được đánh giá 2. Đôi khi chúng ta có thể áp dụng những giá trị biểu diễn một khối số (block of number) hoặc một dải các biến phân loại. Ví dụ, chúng ta có thể sử dụng "1" để biểu diễn giá trị tiền trong khoản từ 0$ đến 20.000$, và sử dụng "2" cho 20.001$ đến 40.000$, và vân vân. Chúng ta có thể sử dụng "0001" để biểu diễn "nhà 2 tầng", và "0002" cho nhà một tầng rưỡi. Tất cả các loại phương pháp "nhanh và tiện" có thể được sử dụng trong chuyển đổi dữ liệu. Không có một thủ tục thống nhất và tiêu chuẩn duy nhất là biến đổi dữ liệu sao cho thuận tiện sử dụng khi khai thác dữ liệu.

Xây dựng mô hình

Xây dựng mô hình dữ liệu là lúc phần mềm khai thác dữ liệu được sử dụng để tạo ra các kết quả. Phân tích nhóm và khảo sát trực quan (visual exploration) thường được áp dụng trước.

Tùy vào dạng của dữ liệu, nhiều mô hình khác nhau có thể được áp dụng. Nếu tác vụ là nhóm dữ liệu lại, và các nhóm được xác định trước thì phân tích biệt số (discriminant analysis) có thể phù hợp. Nếu mục đích là ước lượng thì phân tích hồi qui là phù hợp nếu dữ liệu là dữ liệu liên tục (là phân tích hồi quy logistic nếu không là dữ liệu liên tục). Mạng thần kinh nhân tạo có thể áp dụng cho cả hai trường hợp trên.

Cây quyết định không là một công cụ để phân loại dữ liệu. Những công cụ xây dựng mô hình khác cũng có sẵn. Chúng ta sẽ đi sâu vào những mô hình này hơn trong những chương sau. Điểm quan trọng của khai thác dữ liệu là cho phép người sử dụng làm việc với dữ liệu để đạt được sự hiểu biết. Khai thác dữ liệu thường được khuyến khích sử dụng lặp lại nhiều mô hình.

Xử lý dữ liệu

Khai thác dữ liệu về bản chất là phân tích thống kê, thường sử dụng những bộ dữ liệu lớn. Quy trình chuẩn của khai thác dữ liệu là sử dụng bộ dữ liệu lớn và phân chia nó thành 2 phần, sử dụng một phần của dữ liệu (dữ liệu phân tích) để xây dựng mô hình (bất kể việc sử dụng mô hình nào), và giữ lại một phần của dữ liệu (dữ liệu kiểm tra) để kiểm định mô hình đã xây dựng. Nguyên tắc là nếu xây dựng mô hình dựa trên một bộ dữ liệu cụ thể, đương nhiên mô hình tìm được sẽ có kết quả kiểm tra tốt khi kiểm tra lại mô hình bằng chính bộ dữ liệu dùng để xây dựng mô hình. Bằng việc phân chia dữ liệu và sử dụng một phần của nó để xây dựng mô hình, và kiểm tra mô hình dựa vào bộ dữ liệu được tách ra nên sẽ đạt được tính thuyết phục của kiểm định mô hình.

Ý tưởng về tách dữ liệu thành những thành phần thường mang lại cấp độ cao hơn trong thực hành khai thác dữ liệu. Những phần dữ liệu được phân tách thêm có thể dùng để tinh lọc mô hình.

Các kỹ thuật khai thác dữ liệu

Khai thác dữ liệu có thể thực hiện được bởi các kỹ thuật kết hợp (association), phân lớp (classification), phân cụm (clustering), dự báo (prediction), kiểu mẫu chuỗi (sequential patterns), và chuỗi thời gian tương tự (similar time sequences)[1].

Trong **kỹ thuật kết hợp** (association), mối quan hệ của một mục cụ thể trong dữ liệu giao dịch với những mục khác được sử dụng để dự báo mô hình. Ví dụ: nếu khách hàng mua một máy tính xách tay (X), khi đó anh ta/cô ta sẽ mua một con chuột (Y) với xác suất 60%. Mô hình này xảy ra trong 5,6% của máy tính xách tay bán ra. Theo quy luật kết hợp, tình huống này có thể "X đưa đến Y, trong đó 60% là yếu tố tin cậy (confidence factor) và 5.6% là yếu tố hỗ trợ (support factor)". Khi yếu tố tin cậy và yếu tố hỗ trợ biểu diễn dưới dạng biến ngôn ngữ (linguistic variable) "cao" và "thấp", quy tắc kết hợp có thể được viết dưới dạng logic mờ (fuzzy logic), như "khi yếu tố hỗ trợ là thấp, X dẫn đến Y là cao"[2]. Trong trường hợp có nhiều biến định tính, liên kết mờ (fuzzy association) là cần thiết và là kỹ thuật đầy hứa hẹn trong khai thác dữ liệu.

Trong kỹ thuật **phân lớp** (classification), những phương pháp với chủ đích tìm ra các hàm phân tích khác nhau sắp xếp từng mục của dữ liệu thành một lớp dữ liệu trong một tập các lớp được định nghĩa trước. Đối với một tập các lớp được định nghĩa trước, các thuộc tính, và tập dữ liệu phân tích (learning or training set), phương pháp phân lớp có thể dự báo một cách tự động phân lớp những dữ liệu chưa được phân lớp của tập trong mẫu. Hai vấn đề quan trọng của nghiên cứu liên quan đến kết quả phân lớp là đánh giá phân lớp sai và năng lực dự báo. Các kỹ thuật toán học thường được sử dụng để xây dựng các phương pháp phân lớp như cây quyết định nhị phân, mạng thần kinh, quy hoạch tuyến tính (linear programing) và thống kê. Bằng việc sử dụng cây quyết định nhị phân, mô hình cây quy nạp với định dạng "có-không" có thể được xây dựng để phân tách dữ liệu thành những lớp khác nhau tùy thuộc vào thuộc tính của nó. Mô hình phù hợp với dữ liệu có thể được đo lường bởi hoặc ước lượng thống kê [3] hoặc entropy thông tin [4]. Tuy nhiên, sự phân chia đạt được từ cây qui nạp có thể không tạo ra một giải pháp tối ưu nơi mà năng lực dự báo bị giới hạn. Bằng việc sử dụng mô hình mạng

thần kinh, mô hình qui nạp nơ-ron có thể được xây dựng. Trong cách tiếp cận này, các thuộc tính trở thành các lớp vào (input layers) trong mạng thần kinh khi các lớp kết hợp với dữ liệu là lớp ra (output layers). Giữa lớp vào và lớp ra có một lượng lớn các lớp bị ẩn xử lý độ chính xác của phân lớp. Mặc dù mô hình hồi quy nơ-ron thường đem lại kết quả tốt hơn trong nhiều trường hợp khai thác dữ liệu vì các mối quan hệ đòi hỏi quan hệ phi tuyến phức tạp, triển khai phương pháp này khó khăn khi có nhiều bộ thuộc tính. Trong tiếp cận quy hoạch tuyến tính, vấn đề phân lớp được xem như một dạng đặc biệt của quy hoạch tuyến tính [5]. Khi cho sẵn một bộ các lớp và một bộ thuộc tính của các biến, người ta có thể định nghĩa các giới hạn phân chia các lớp. Khi đó mỗi lớp được biểu diễn bởi một nhóm các ràng buộc đối với các giới hạn trong quy hoạch tuyến tính. Hàm mục tiêu trong mô hình quy hoạch tuyến tính có thể giảm tối thiểu tỷ lệ trùng lặp chéo giữa các lớp và tối đa khoảng cách giữa các lớp[6]. Cách tiếp cận quy hoạch tuyến tính đưa đến kết quả phân lớp tối ưu. Nó cũng rất khả thi để xây dựng một sự phân cách hữu hiệu trong hiện tượng đa lớp. Tuy nhiên, thời gian tính toán có thể vượt quá thời gian nếu dùng cách tiếp cận thống kê. Nhiều phương pháp thống kê như hồi quy tuyến tính sai biệt, hồi quy toàn phương sai biệt và, hồi quy logistic sai biệt là rất thông dụng và được sử dụng rộng rãi trong phân lớp kinh doanh thực. Mặc dù phần mềm thống kê được xây dựng để xử lý những khối dữ liệu lớn, các tiếp cận thống kê có một bất lợi trong phân cách hữu hiệu hiện tượng đa lớp trong đó so sánh cặp (so sánh một lớp với tất cả các lớp còn lại) được chấp nhận.

Phân tích phân nhóm (clustering) dùng dữ liệu chưa được phân theo nhóm và sử dụng những kĩ thuật tự động để phân dữ liệu này vào các nhóm. Phân nhóm là phương pháp không có tính giám sát, là không đòi hỏi một bộ dữ liệu trong mẫu. Nó chia sẻ một phương pháp luận nền tảng chung với "phân lớp" (classification). Nói một cách khác, hầu hết các mô hình toán học đã nhắc đến trước đầu liên quan đến "phân lớp" đều có thể áp dụng cho phân tích phân nhóm. Phân tích phân nhóm sẽ được mô tả trong chương 5.

Phân tích dự báo (prediction analysis) liên quan đến các kỹ thuật hồi qui. Ý tưởng quan trọng của phân tích dư báo là khám phá ra các mối quan hệ giữa các biến phụ thuộc và biến độc lập, mối quan hệ giữa các biến độc lập (một biến đối với một biến khác, một biến đối với các biến còn lại, vv…). Ví dụ, nếu doanh thu là biến độc lập, thì lợi nhuận có thể là biến phụ thuộc. Bằng các sử dụng dữ liệu quá khứ của doanh thu và lợi nhuận, hoặc dùng kĩ thuật hồi quy tuyến tính hoặc hồi quy phi tuyến để xây dựng một đường cong hồi quy được sử dụng để dự báo lợi nhuận trong tương lai.

Mô hình phân tích chuỗi (sequential pattern analysis) tìm các mô hình tương đồng trong các thao tác/giao dịch dữ liệu qua các thời kỳ kinh doanh. Những mô hình này có thể được sử dụng bởi các chuyên viên phân tích kinh doanh để nhận dạng mối quan hệ giữa các dữ liệu. Những mô hình toán học đằng sau phân tích chuỗi là nguyên tắc logic, logic mờ, và vv…Như một sự mở rộng của mô hình phân tích chuỗi, phân tích chuỗi thời gian tương tự (similar time sequences) được ứng dụng để phát hiện các chuỗi tương đồng với chuỗi đã biết ở các thời kỳ kinh doanh trong quá khứ và hiện tại. Trong khai thác dữ liệu, vài chuỗi tương tự có thể được nghiên cứu để nhận định xu hướng tương lai của phát triển giao dịch. Các tiếp cận này hữu ích trong việc xử lý các cơ sở dữ liệu theo chuỗi thời gian.

Đánh giá

Giai đoạn diễn dịch dữ liệu rất quan trọng. Nó giúp tiêu hóa kiến thức từ dữ liệu khai thác được. Có 2 vấn đề cốt yếu. Một là làm thế nào nhận ra giá trị kinh doanh từ những mô hình tri thức phát hiện ra trong giai đoạn khai thác dữ liệu. Hai là công cụ trực quan nào nên sử dụng để trình bày các kết quả khai thác dữ liệu. Xác định giá trị kinh doanh từ các mô hình tri thức được phát hiện cũng tương tự như chơi ô chữ (puzzles). Các dữ liệu khai thác được là một ô chữ cần được lắp ghép lại để phục vụ cho mục tiêu kinh doanh. Hoạt động này phụ thuộc vào sự tương tác giữa các chuyên viên phân tích dữ liệu, chuyên viên phân tích kinh doanh và những người ra quyết định (như các nhà quản lý, giám đốc điều hành). Bởi vì những chuyên

viên phân tích dữ liệu có thể không nhận thức đầy đủ mục đích của khai thác dữ liệu hoặc mục tiêu kinh doanh, trong khi đó các chuyên viên phân tích kinh doanh không hiểu được các kết quả của các giải pháp toán học phức tạp, họ tương tác với nhau là cần thiết. Để diễn giải một cách hợp lý mô hình tri thức, điều quan trọng là phải chọn được công cụ trực quan thích hợp. Có nhiều công cụ trực quan hiện hành như biểu đồ hình tròn (pie chart), biểu đồ phân phối tần số (histograms), biểu đồ hộp (box plot), đồ thị phân tán (scatter plots) và các phân phối (distributions). Sự diễn dịch tốt sẽ dẫn đến những quyết định kinh doanh hữu hiệu, trong khi những diễn dịch nghèo nàn có thể bỏ sót các thông tin có ích. Thông thường, diễn dịch bằng đồ thị càng đơn giản người sử dụng càng dễ hiểu.

Triển khai

Kết quả của nghiên cứu khai thác dữ liệu cần được báo cáo cho người đặt hàng dự án. Nghiên cứu khai thác dữ liệu khám phá những kiến thức mới mà những kiến thức này cần được gắn chặt với các mục đích của dự án khai thác dữ liệu. Ban quản lý dự án sẽ là người ứng dụng những tri thức mới này vào môi trường kinh doanh của họ.

Quan trọng là kiến thức đạt được từ những nghiên cứu cụ thể về khai thác dữ liệu phải được giám sát vì sự thay đổi. Hành vi khách hàng thay đổi theo thời gian, và những điều đúng trong thời gian thu thập dữ liệu có thể đã bị thay đổi. Nếu những thay đổi cơ bản xuất hiện, tri thức đạt được sẽ không còn đúng nữa. Vì vậy, điều quan trọng là lĩnh vực quan tâm phải được giám sát trong thời gian triển khai.

QUY TRÌNH KHÁM PHÁ TRI THỨC

Một nghiên cứu khai thác dữ liệu gần đây trong lĩnh vực bảo hiểm đã ứng dụng quy trình khám phá tri thức[7]. Quy trình này áp dụng lặp lại các bước đã nhắc đến trong CRISP-DM, và thể hiện phương pháp luận được thực hiện như thế nào trong thực tiễn.

Giai đoạn 1. Hiểu yêu cầu ứng dụng trong kinh doanh: một mô hình cần để dự báo sớm khách hàng nào sẽ mất khả năng thanh toán giúp cho hãng đưa ra các biện pháp ngăn chặn (hoặc biện pháp để ngăn ngừa mất những khách hàng tốt). Mục đích là tối thiểu hóa việc phân lớp sai các khách hàng hợp lệ.

Trong trường hợp này thời gian thanh toán hóa đơn điện thoại là 2 tháng. Khách hàng sử dụng điện thoại trong 4 tuần, và nhận hóa đơn sau một tuần. Thanh toán đến hạn một tháng sau ngày xuất gởi hóa đơn. Thường các hãng cho khách hàng khoảng 2 tuần sau ngày đến hạn thanh toán trước khi họ hành động, vào thời gian này điện thoại sẽ bị ngắt kết nối nếu số tiền chưa thanh toán lớn hơn một mức nào đó. Hóa đơn sẽ tiếp tục được gởi mỗi tháng trong 6 tháng, trong thời gian này khách hàng nên dàn xếp việc thanh toán. Nếu không nhận được khoảng thanh toán trong vòng 6 tháng, khoản chưa thanh toán sẽ được chuyển sang phân loại không thu được tiền.

Trường hợp này giả định khách hàng mất khả năng thanh toán có thể thay đổi thói quen gọi điện và việc sử dụng điện thoại trong thời gian trước khi và ngay sau khi kết thúc thời hạn thanh toán hóa đơn. Thay đổi thói quen gọi điện, kết hợp với mô hình thanh toán được kiểm định khả năng cung cấp dự báo tốt về tình trạng mất khả năng thanh toán trong tương lai.

Giai đoạn 2. Hiểu dữ liệu: thông tin tĩnh của khách hàng có sẵn trong hồ sơ khách hàng. Dữ liệu theo thời gian có sẵn trên hóa đơn, việc thanh toán, và sử dụng dịch vụ. Dữ liệu từ một số cơ sở dữ liệu nhưng các cơ sở dữ liệu này là thuộc nội bộ của công ty. Kho dữ liệu được xây dựng để tập hợp và tổ chức các dữ liệu này. Dữ liệu được mã hóa để bảo vệ bí mật của khách hàng. Dữ liệu bao gồm thông tin khách hàng, việc sử dụng điện thoại từ các tổng đài trung tâm, thông tin thanh toán, báo cáo thanh toán của khách hàng, việc tạm ngừng dịch vụ điện thoại do không thanh toán, dịch vụ điện thoại được kết nối lại sau khi thanh toán, và báo cáo về hủy bỏ hợp đồng vĩnh viễn.

Dữ liệu được thu thập cho 100.000 khách hàng trong 17 tháng, và được thu thập từ một vùng nông thôn/nông nghiệp, một vùng bán nông nghiệp, và một vùng đô thị/công nghiệp để bảo đảm tính đại diện cho cơ sở dữ liệu về khách hàng của hãng. Kho dữ liệu sử dụng hơn 10 gigabytes để lưu trữ dữ liệu thô.

Giai đoạn 3. Chuẩn bị dữ liệu: Dữ liệu sẽ được kiểm tra về chất lượng, và dữ liệu không có ích cho nghiên cứu sẽ bị lọc ra. Những dữ liệu hỗn tạp thường có tương quan với nhau. Ví dụ, rõ ràng những cuộc gọi cước ít có ít ảnh hưởng đến nghiên cứu. Điều này cho phép giảm được 50% khối lượng dữ liệu. Tỷ lệ các trường hợp gian lận càng thấp làm cho việc làm sạch dữ liệu càng cần thiết khỏi những giá trị bỏ sót hoặc sai lầm do những cách ghi nhận dữ liệu khác nhau trong tổ chức và sự phân tán của nguồn dữ liệu. Vì vậy cần thiết phải kiểm tra chéo dữ liệu như ngắt kết nối điện thoại. Dữ liệu trễ đòi hỏi sự đồng bộ của các yếu tố dữ liệu khác nhau.

Đồng bộ dữ liệu sẽ lộ ra số khách hàng không có khả năng chi trả với những thông tin bị bỏ sót đã bị xóa từ bộ dữ liệu. Theo cách đó cần giảm và dự phóng dữ liệu, như thế thông tin được nhóm theo tài khoản khiến dữ liệu được dùng cho việc thao tác một cách dễ dàng hơn, và dữ liệu khách hàng được gộp thành thời đoạn 2 tuần. Thống kê được áp dụng để tìm ra đặc tính thể hiện yếu tố phân biệt (discriminant factor) giữa các khách có khả năng và không có khả năng thanh toán. Dữ liệu gồm như sau:

- Phân loại tài khoản điện thoại (23 loại, như điện thoải trả tiền (payphone), kinh doanh, vv…)
- Khoản nợ trung bình được tính toán cho tất các khách hàng có khả năng và không có khả năng trả nợ. Khách hàng không có khả năng trả nợ có dư nợ trung bình cao hơn hẳn trong tất cả các loại tài khoản.
- Khoản phụ thu trên hóa đơn được xác định bằng cách so sánh tổng số tiền phải trả cho việc sử dụng điện thoại trong khoảng thời gian khảo sát với số dư trả trước hoặc trả tiền do mua các thiết bị phần cứng hoặc các dịch vụ khác. Điều này cũng chứng minh rằng có sự khác biệt có ý nghĩa thống kê giữa 2 loại khách hàng kết quả.
- Thanh toán trả góp cũng được khảo sát. Tuy nhiên, biến này không có ý nghĩa thống kê.

*Giai đoạn 4. Xây dựng mô hình:*Vấn đề của dự báo là phân lớp, với 2 lớp: có thể thanh toán (99,3% các trường hợp) và hầu như không thể thanh toán (0,7% các trường hợp). Vì vậy, việc đếm các trường hợp không có khả năng thanh toán là rất nhỏ trong một thời đoạn thanh toán. Chi phí của các sai sót biến thiên nhiều trong 2 phân lớp. Điều này được ghi lại bởi nhiều người như một vấn đề khó khăn trong phân lớp.

Một bộ dữ liệu mới được tạo nên thông qua chọn mẫu phân tầng các khách hàng có khả năng thanh toán, thay đổi phân phối của khách hàng thành 90% có khả năng trả nợ và 10% không có khả năng trả nợ. Tất cả các trường hợp không trả nợ được giữ lại, và cẩn thận duy trì tỷ phần tương ứng theo vùng địa lý, theo loại điện thoại kết nối, theo các nhóm tài khoản điện thoại cho bộ dữ liệu khách hàng thanh toán được. Một bộ dữ liệu 2.066 trường hợp được xây dựng.

Thời gian khảo sát cho mỗi tài khoản điện thoại sẽ được thiếp lập. Đối với những tài khoản bị hủy bỏ, thời gian khảo sát này là thời gian 15 đoạn thời gian 2 tuần cuối trước khi dịch vụ bị cắt. Đối với những tài khoản vẫn hoạt động, thời gian khảo sát được chọn như khoản thời gian tương tự như trường hợp ngừng dịch vụ. Có 6 ngày có khả năng là ngừng dịch vụ mỗi năm. Đối với những tài khoản đang hoạt động, một trong 6 ngày này được chọn một cách ngẫu nhiên.

Đối với mỗi khách hàng, các biến được định nghĩa bằng các phép đo lường thích hợp cho mỗi thời đoạn 2 tuần trong thời gian khảo sát của quan sát đó. Tại cuối giai đoạn này, các biến mới được tạo ra để mô tả việc sử dụng điện thoại của khách hàng so sánh với trung bình di động

(moving average) của 4 thời đoạn 2 tuần trước đó. Tại giai đoạn này, có 46 biến có thể dùng làm nhân tố phân biệt. Những biến này bao gồm 40 biến được tính toán như thói quen gọi điện trong 15 thời đoạn 2 tuần, cũng như các biến liên quan đến loại khách hàng, khách hàng là mới hay cũ, và 4 biến liên quan đến việc thanh toán hóa đơn.

Các thuật toán phân tích phân biệt, cây quyết định và mạng thần kinh được sử dụng để kiểm định giả thuyết đối với bộ dữ liệu rút gọn 2.066 trường hợp xử lý trên 46 biến.

Giai đoạn 5. Đánh giá: Thử nghiệm được tiến hành để kiểm tra và so sánh kết quả thực hiện. Dữ liệu được chia thành bộ phân tích (khoảng hai phần ba của 2.066 trường hợp) và bộ kiểm tra (dữ liệu còn lại). Sai sót phân lớp thường xuất hiện trong ma trận trùng (coincidence matrix) (còn được một số người gọi là ma trận nhầm - confusion matrix). Ma trận trùng biểu thị số trường hợp được phân lớp đúng, cũng như số trường hợp phân lớp không đúng trong mỗi bộ dữ liệu (category). Nhưng trong nhiều nghiên cứu khai thác dữ liệu, mô hình có thể rất tốt trong phân lớp ở một loại dữ liệu này, trong khi rất tệ khi phân lớp ở các loại dữ liệu khác. Giá trị chủ yếu của ma trận trùng là nhận ra loại lỗi đã mắc phải. Điều quan trọng hơn nhiều là tránh một loại lỗi nào đó hơn các loại khác. Ví dụ, nhân viên tín dụng của ngân hàng gánh phải hậu quả việc cho một ai đó vay, người này được kì vọng sẽ trả tiền nhưng lại không trả, mà mắc phải một sai lầm tiếp là không cho những những người khác vay, những người này lại thực sự sẽ trả nợ. Cả hai trường hợp đều là lỗi phân lớp, nhưng trong khai thác dữ liệu, thường thì một loại lỗi nào đó sẽ quan trọng hơn rất nhiều so với các loại còn lại. Ma trận trùng cung cấp một phương thức để tập trung vào các loại lỗi cụ thể mà mô hình có khuynh hướng mắc phải.

Một cách để phản ảnh các lỗi quan trọng là thông qua chi phí. Đây là một ý tưởng tương đối đơn giản, cho phép người sử dụng tính toán chi phí tương đối theo loại lỗi. Ví dụ, nếu mô hình dự báo một tài khoản mất khả năng thanh toán, điều này có nghĩa là mất đi một khoản tiền nợ trung bình 200$. Mặt khác, chi phí phát sinh khi chờ một tài khoản mà rốt cuộc sẽ được trả có thể chỉ tốn 10$. Vì vậy, có thể có một khác biệt quan trọng trong chi phí của sai lầm trong trường hợp này. Xử lý một trường hợp thanh toán được như một tài khoản chết có nguy cơ mất mát 190$, bên cạnh đó gia tăng xa lánh với khách hàng (những người mà có thể hoặc không thể có lợi trong tương lai), Ngược lại, xử lý tài khoản sẽ không bao giờ thanh toán bằng cách giữ tài khoản này trong sổ sách lâu hơn cần thiết sẽ tốn thêm chi phí 10$. Đến đây, một hàm chi phí cho ma trận trùng là:

190$ x (đóng tài khoản tốt) + 10$ x (giữ các tài khoản xấu)

(Lưu ý rằng chúng ta sử dụng đồng đôla để ví dụ, không phải là số thực tế). Phép đo lường này (giống như tỷ lệ phân lớp đúng – correct classification rate) có thể được sử dụng để so sánh các mô hình khác.

SPSS được sử dụng cho **phân tích phân biệt**, bao gồm một thủ tục đưa vào từng bước (stepwise forward selection procedure). Mô hình tốt nhất gồm 17 trong số 46 biến. Sử dụng chi phí phân lớp sai ngang bằng (equal misclasscification cost) mang lại ma trận trùng trong Bảng 2.1.

Bảng 2- 1: Ma trận trùng – Chi phí phân lớp ngang bằng

Hóa đơn điện thoại	Mất khả năng thanh toán theo mô hình	Có khả năng thanh toán theo mô hình	
Thực sự mất khả năng thanh toán	50	14	64
Thực sự có khả năng thanh toán	76	578	654
	126	592	718

Độ chính xác phân lớp chung đạt được bằng cách phân chia số phân lớp chính xác (50+578=628) bằng tổng số các trường hợp (718). Vì vậy, kiểm tra dữ liệu được phân lớp đúng là 87.5% của các trường hợp. Hàm chi phí trong trường hợp này là:

190$ x 76 + 10$ x 14 =14.580$

Bảng 2- 2: Ma trận trùng – Chi phí phân lớp không ngang bằng

Hóa đơn điện thoại	Mất khả năng thanh toán theo mô hình	Có khả năng thanh toán theo mô hình	
Thực sự mất khả năng thanh toán	36	28	64
Thực sự có khả năng thanh toán	22	632	654
	58	660	718

Tỷ lệ cao của các trường hợp thực sự có khả năng chi trả được phân lớp như không có khả năng chi trả được xem là không chấp nhận được, bởi vì nó xua đuổi nhiều khách hàng tốt. Tiến hành lại một thực nghiệm sử dụng xác suất ưu tiên. Điều này tăng độ tin cậy đầu ra, được biểu thị trọng ma trận trùng ở Bảng 2.2.

Vì vậy, dữ liệu kiểm tra được phân lớp một cách đúng đắn trong 93% của các trường hợp. Đối với dữ liệu phân tích, tính toán này là 93.6%. Mô hình thường khớp với dữ liệu phân tích tốt hơn một chút so với dữ liệu kiểm tra, bởi vì chúng được xây dựng trên chính dữ liệu phân tích. Dữ liệu kiểm tra độc lập giúp kiểm tra tốt hơn nhiều. Độ chính xác của khác hàng mất khả năng trả nợ là quan trọng hơn rất nhiều vì nó tốn kém nhiều hơn rất nhiều, giảm từ 78% trong dữ liệu phân tích còn 56% trong dữ liệu kiểm tra. Hàm chi phí như sau:

190$ x 22 + 10$ x 28 = 4.460$

Xét từ tổng chi phí, mô hình sử dụng phương pháp chi phí phân lớp không ngang bằng (sử dụng chi phí thực) được xem hữu ích hơn.

17 biến được xác định trong phân tích phân biệt được dùng để sử dụng thêm cho hai mô hình khác. Bộ dữ liệu phân tích và kiểm tra tương tự được sử dụng. Dữ liệu phân tích dùng để xây dựng một mô hình phân loại nguyên tắc. Ma trận trùng cho bộ dữ liệu kiểm tra được biểu diễn ở Bảng 2.3.

Bảng 2- 3: Ma trận trùng – Mô hình nguyên tắc

Hóa đơn điện thoại	Mất khả năng thanh toán theo mô hình	Có khả năng thanh toán theo mô hình	
Thực sự mất khả năng thanh toán	38	26	64
Thực sự có khả năng thanh toán	8	646	654
	46	672	718

Dữ liệu kiểm tra phân loại đúng 95.26% các trường hợp. Đối với dữ liệu phân tích, tính toán là 95.3%. Hàm chi phí là:

190$ x 8 + 10$ x 26 =1.780$

Đây là một bước tiến trong mô hình phân tích phân biệt.

Một số thực nghiệm được tiến hành với **mô hình mạng thần kinh** sử dụng 17 biến và dữ liệu trong mẫu tương tự. Kết quả ma trận trùng đối với dữ liệu kiểm tra được thể hiện ở Bảng 2.4.

Dữ liệu kiểm tra được phân lớp đúng với 92.9% các trường hợp. Đối với dữ liệu phân tích, tính toán được 94.1%. Hàm chi phí như sau:

190$ x 11 + 10$ x 40 = 2.490$

Tuy nhiên, những kết quả này kém hơn kết quả tính từ mô hình quyết định.

Bảng 2- 4: Ma trận trùng – Mô hình mạng thần kinh.

Hóa đơn điện thoại	Mất khả năng thanh toán theo mô hình	Có khả năng thanh toán theo mô hình	
Thực sự mất khả năng thanh toán	24	40	64
Thực sự có khả năng thanh toán	11	643	654
	35	683	718

Mục tiêu đầu tiên là tối đa độ chính xác của việc dự báo các khách hàng mất khả năng thanh toán. Dường như cây quyết định thực hiện việc này tốt nhất. Mục tiêu thứ hai là tối thiểu tỷ lệ

lỗi đối với những khách hàng có khả năng trả nợ. Mô hình mạng thần kinh cho kết quả gần nhất với kết quả của mô hình cây quyết định. Cả 3 mô hình này đều được quyết định sử dụng tùy từng trường hợp.

Giai đoạn 6. Triển khai: Mỗi khách hàng được xem xét sử dụng cả 3 thuật toán. Nếu cả 3 đều có kết quả phân lớp thống nhất, kết quả đó sẽ được sử dụng. Nếu các kết quả của mô hình không thống nhất thì khách hàng được xác định là không phân lớp được. Sử dụng kiểu này cho bộ dữ liệu kiểm tra đạt được ma trận trùng như trong bảng 2.5.

Bảng 2- 5: Ma trận trùng – Mô hình kết hợp

Hóa đơn điện thoại	Mất khả năng thanh toán theo mô hình	Có khả năng thanh toán theo mô hình	Không phân loại	
Thực sự mất khả năng thanh toán	19	17	28	64
Thực sự có khả năng thanh toán	1	626	27	654
	20	643	55	718

Dữ liệu kiểm tra phân lớp đúng được 89.8% các trường hợp. Nhưng chỉ một khách hàng thực sự có khả năng thanh toán đã bị ngắt kết nối mà không được phân tích thêm. Hàm chi phí là:

190$ x 1 + 10$ x 17 = 360$.

Những bước sử dụng trong ứng dụng này hợp với 6 giai đoạn đã được trình bày. **Chọn lọc dữ liệu** liên quan đến việc hiểu biết các lĩnh vực ứng dụng và tạo bộ dữ liệu mục tiêu. **Tiền xử lý dữ liệu** liên quan đến việc **làm sạch dữ liệu** và tiền xử lý. **Chuyển đổi dữ liệu** liên quan đến **việc giảm dữ liệu** và dự phóng. Khai thác dữ liệu được mở rộng trong những ứng dụng trước đây gồm: (1) lựa chọn hàm cho khai thác dữ liệu, (2) chọn thuật toán khai thác dữ liệu, (3) khai thác dữ liệu. **Diễn dịch dữ liệu** bao gồm diễn dịch và sử dụng tri thức khám phá được.

TÓM TẮT

Quy trình khai thác dữ liệu chuẩn CRISP-DM có 6 giai đoạn: (1) Hiểu yêu cầu, (2) Hiểu dữ liệu, (3) Chuẩn bị dữ liệu, (4) Xây dựng mô hình, (5) Đánh giá và (6) triển khai. Chọn lọc và hiểu dữ liệu, chuẩn bị và diễn dịch mô hình yêu cầu sự làm việc nhóm giữa các chuyên viên phân tích khai thác dữ liệu và chuyên viên phân tích kinh doanh, trong khi đó chuyển đổi dữ liệu và khai thác dữ liệu được tiến hành bởi các chuyên viên khai thác dữ liệu một cách riêng biệt. Mỗi giai đoạn là một bước chuẩn bị cho giai đoạn tiếp theo. Trong những chương còn lại của quyển sách này, chúng ta sẽ thảo luận chi tiết quy trình này theo những góc độ khác nhau, như công cụ khai thác dữ liệu và ứng dụng. Điều này sẽ cung cấp cho người đọc hiểu rõ hơn tại sao một quy trình đúng đắn còn quan trọng hơn so với thực hiện đúng phương pháp.

Chú thích thuật ngữ

Association (kết hợp): chức năng khai thác dữ liệu nhằm nhận dạng những mô hình/kiểu tương quan.

Classification (phân lớp): phân tích để sắp xếp các trường hợp/quan sát vào những phân lớp khác nhau

Clustering (nhóm): phân tích nhằm nhóm các dữ liệu vào các lớp

Coincidence matrix (ma trận trùng): bảng trình bày các tần số quan sát thực tế và các tần số theo mô hình dự báo.

Correct classification rate (tỷ lệ phân lớp đúng): tỷ lệ những trường hợp được phân lớp đúng so với tổng số trường hợp trong mẫu kiểm tra.

Cost function: tổng số trường hợp kiểm tra sai nhân với chi phí ước tính của từng loại lỗi.

Demographic data (dữ liệu nhân khẩu): dữ liệu liên quan đến đặc tính dân số

Imputation: điền những thông tin bị khuyết với trị số thích hợp dựa vào thông tin lân cận.

Prediction analysis: phân tích mối liên hệ giữa trị số của các trường hợp với trị số của các biến giải thích

Qualitative data (dữ liệu định tính): dữ liệu không đo được bằng số

Quantitative data (dữ liệu định lượng): dữ liệu đo được bằng số

Sequential pattern analysis: tìm kiếm các mô hình tương đồng

Similar time sequenses: tìm kiếm các trình tự trong dữ liệu

Socio-graphic data: dữ liệu liên quan đến các hoạt động văn hóa

Test set (tập kiểm tra): một phần của dữ liệu hiện hành được dùng để kiểm tra các mô hình khai thác dữ liệu

Training set (tập phân tích): một phần của dữ liệu hiện hành dùng để xây dựng mô hình khai thác dữ liệu

Transactional data (dữ liệu giao dịch): dữ liệu liên quan đến hoạt động kinh doanh ở mức độ cơ bản

Bài tập

1. Thảo luận 3 vấn đề liên quan đến việc lựa chọn dữ liệu cho phân tích khai thác dữ liệu.
2. So sánh những điểm khác biệt về dữ liệu nhân khẩu (dân số), dữ liệu xã hội, và dữ liệu giao dịch.
3. Khác nhau giữa dữ liệu định lượng và dữ liệu định tính.
4. Dữ liệu thừa là gì?
5. Tại sao một phần mềm khai thác dữ liệu không thể hỗ trợ cho tất cả các phân tích khai thác dữ liệu?
6. Phần mềm khai thác dữ liệu có thể loại bỏ nhu cầu người sử dụng cần hiểu các phương pháp thống kê?
7. Có phải dạng phân tích thống kê tốt để kiểm tra độ chính xác của mô hình trên tập dữ liệu được sử dụng để xây dựng mô hình?
8. Chuyển đổi những dữ liệu số sau đây liên quan đến tuổi của khách hàng thành dạng trẻ (nếu tuổi dưới 40), trung niên (từ 40 đến 60 tuổi), và già (lớn hơn 60 tuổi)

Khách hàng	Tuổi
Fed	46
Herman	52
George	36
Frieda	39
Hermione	28

9. Chuyển đổi mức lương dưới đây thành thang đo số với 20.000$ bằng 0, 220.000$ bằng 1.0, và tất cả các mức khác thành các trị số tương ứng tuyến tính giữa 0 và 1.

Khách hàng	Lương
Fed	120.000$
Herman	200.000$
George	50.000$
Frieda	65.000$
Hermione	35.000$

10. Tại sao giá trị danh nghĩa không thể chuyển đổi sang giá trị có thang đo số?

11. Với dữ liệu được cho sau đây, những mô hình khai thác dữ liệu khác nhau được áp dụng trong việc kiểm định bộ dữ liệu. Mô hình nào tốt nhất?

Đối tượng	Thực	Hồi quy	Phân nhóm	Mạng thần kinh	Cây quyết định
Fed	Tốt	Tốt	Tốt	Tốt	Tốt
Herman	Xấu	Tốt	Xấu	Tốt	Xấu
George	Tốt	Tốt	Xấu	Xấu	Xấu
Frieda	Tốt	Tốt	Tốt	Tốt	Xấu
Hermione	Xấu	Tốt	Xấu	Xấu	Tốt

12. Lập ma trận trùng cho mỗi mô hình trong 4 mô hình ở câu 11.

13. Trong câu 12, mô hình hồi quy và mạng thần kinh có số lỗi như nhau. Tại sao kết quả của một mô hình được đánh giá tốt hơn kết quả của các mô hình khác.

14. Trong câu 11, cho biết chi phí lỗi tương đối của kết quả thực sự tốt bị phân lớp thành xấu là 100$ và chi phí lỗi tương đối của kết quả thực sự xấu bị phân lớp thành tốt là 500$, hãy xác định tổng chi phí lỗi cho mỗi mô hình trong 4 mô hình trên.

Kết chú

[1] P.Cabena, P.Hadjinia, R.Stadler, J.Verhees, vaf A.Zanasi, *Discovering data Mining from Concepts to Implementation* (Upper Sadle River, NJ: nhà xuất bản Prentice Hall, 1997).

[2] J.-S.R.Jang, C.-T.Sun, and E.Mizutani, *Neuro-Fuzzy and Soft Computating* (Upper Sadle River, NJ: nhà xuất bản Prentice Hall, 1997).

[3] L.Breiman, J.Friedman, R.Olshen, vaf C.Stone, *Classification and Regression Trees* (Belmont: Wadsworth, 1984).

[4] J.Quilan, "Induction of decisioon trees", *Machine learning,* cuốn 1 (1986), trang 81-106.

[5] N.Freed và F.Glover,"Simple but Powerful Goal Programming Models for Discriminant Problems", *European Journal of Operational Research,* cuốn 7 (1981), trang 44-60; Y.Shi và P.L.Yu, "Goal Setting and Compromise Solutions", chỉnh sửa bởi B.Karpak và S.Zionts (Berling: nhà xuất bản Springer-Verlag, 1989), trang 165-204.

[6] Y.Shi, "Multiple Criteria Decision Making in Credit Card Portfolio", (Khoa công nghệ thông tin, Đại học Nebraska ở Omaha, 1998).

[7] S.Daskalaki, I. Kopanas, M.Goudara, vaf A.Avouris, "Data Mining for Decision Support on Customer Insolvency in the Telecommunications Business", European Jounal of Operational Research, cuốn 145 (2003), trang 239-255.