METODY STOSOWANE W DATA MINING

prof. dr hab. Andrzej Sokołowski, Akademia Ekonomiczna w Krakowie, Katedra Statystyki

W procesie data mining, po przygotowaniu danych, przychodzi etap właściwego przeszukiwania, przerzucania danych, zgłębiania ich, kopania w pokładach informacji. Poszukujemy ukrytych prawidłowości lub anomalii, poszukujemy wiedzy w ogromie danych. Przedstawienie uporządkowanego przeglądu metod wówczas stosowanych jest zadaniem niesłychanie trudnym. Z jednej strony wykorzystuje się metody znane z klasycznej statystyki (często bardzo proste), także niekonwencjonalne metody eksploracji danych, a z drugiej zaproponowano wiele nowych procedur, które obficie korzystają z rozwoju możliwości obliczeniowych współczesnych komputerów. Można stosować algorytmy wymagające nawet bardzo skomplikowanych obliczeń, a „liczba" danych, które wykorzystujemy w data mining, może być niemal dowolnie duża.

W literaturze przedmiotu brak powszechnie uznanej typologii metod. Przytaczamy tu trzy różne podejścia.

Weiss i Indurkhya [1998] stwierdzają, że metody prognozowania w podejściu data mining dostarczają trzech typów rozwiązań: matematycznych, dystansowych oraz logicznych. Rozwiązania matematyczne i logiczne wykorzystują operacje na wartościach cech charakteryzujących nowy obiekt, zaś rozwiązanie dystansowe korzysta z informacji o obiektach leżących blisko obiektu analizowanego.

Wspomniani autorzy wyróżniają trzy typy rozwiązań matematycznych (a więc takich, które wykorzystują łączną informację z różnych cech z wykorzystaniem prostych operacji arytmetycznych):

- liniowe – dyskryminacja liniowa, regresja liniowa,
- sieci neuronowe,
- zaawansowane metody statystyczne – *projection pursuit*, MARS (*Multiple Adaptive Regresion by Splines – Wieloraka adaptacyjna regresja składana*).

Rozwiązanie dystansowe polega na znalezieniu *k* najbliższych sąsiadów dla nowego obiektu i wnioskowaniu o brakujących cechach tego obiektu na podstawie analogicznych cech jego najbliższych sąsiadów. Ważną decyzją jest tu wybór miary odległości oraz wybór liczby sąsiadów, których mamy wziąć pod uwagę. To, czy dobór liczby sąsiadów jest odpowiedni, można sprawdzić, obserwując, jak zmienia się błąd klasyfikacji (lub predykcji) w miarę wzrostu *k*.

Metody logiczne wykorzystują informacje z próby, łącząc je poprzez operacje logiczne *Prawda/Fałsz*. Najbardziej znane procedury z tej grupy to *Drzewa decyzyjne* oraz *Zasady decyzyjne*.

Berry i Linoff [2000] wyróżniają trzy techniki data mining:

- automatyczne wykrywanie skupień,
- drzewa decyzyjne,
- sieci neuronowe.

Z kolei Hastie, Tibshirani i Friedman [2001] dzielą metody na *supervised* i *unsupervised*. Wydaje się, że przyjęte u nas tłumaczenie tych terminów jako metody „z nauczycielem" i „bez nauczyciela" nie pokrywa się z ideą autorów książki. Proponujemy termin *ukierunkowane* dla metod (problemów), których celem jest zaprognozowanie wartości zmiennej wyjściowej na podstawie realizacji zmiennych wejściowych. W metodach *nieukierunkowanych* celem jest znalezienie związków i prawidłowości (wzorców) na podstawie tylko cech wejściowych. Do metod ukierunkowanych zaliczają bardzo szeroko rozumianą regresję, wnioskowanie poprzez znajdowanie *k* najbliższych sąsiadów i sieci neuronowe. Metody *nieukierunkowane* obejmują badanie zasad asocjacji, analizę skupień, samoorganizujące się mapy oraz metodę głównych składowych.

Jak widać, różne są koncepcje typologii metod data mining. Zdecydowaliśmy się przedstawić podstawy i główne idee czterech grup metod, zgodnie zaliczanych do metod data mining przez wszystkich autorów zajmujących się ta problematyką.

Drzewa decyzyjne

Gdy w celu podzielenia pewnego zbioru obiektów dokonujemy kolejnych podziałów zakresu zmienności cech statystycznych opisujących obiekty, to proces ten wygodnie jest przedstawić w postaci drzewa. Jego hierarchiczna struktura obrazuje proces podejmowania decyzji. Wśród cech opisujących nasze obiekty jest jedna szczególnie wyróżniona. W zależności od typu rozważanego zagadnienia jest ona nazywana: zmienną objaśnianą, zależną, prognozowaną, definiującą przynależność. Jeżeli zmienna ta ma charakter ilościowy, to drzewo zbudowane dla wyjaśnienia jej kształtowania się nazywane jest *drzewem regresyjnym*. Jeżeli zmienna objaśniana jest zmienną jakościową – określającą najczęściej przynależność do konkretnej klasy obiektów - to wtedy mamy do czynienia z *drzewem klasyfikacyjnym*.

Metodologię drzew decyzyjnych przedstawimy na przykładzie drzew klasyfikacyjnych. Formalnie rzecz biorąc, drzewo jest grafem składającym się z wierzchołków i krawędzi łączących niektóre wierzchołki. Najprostsze drzewa to tzw. drzewa binarne, w których z każdego wierzchołka wychodzą dwie krawędzie. Każdy taki wierzchołek reprezentuje decyzję o podziale zbioru (lub podzbioru) obiektów na dwa podzbiory ze względu na jedną z cech objaśniających. Początkowy wierzchołek drzewa, obrazujący pierwszą decyzję

podziału, nazywamy *korzeniem drzewa*. Z kolei *liściem drzewa* nazywamy wierzchołek, z którego nie wychodzą żadne krawędzie. Na tym etapie następuje identyfikacja obiektu.

Podstawowa metodologia konstrukcji drzewa opiera się na zasadzie *rekurencyjnego podziału*. Zasadnicze znaczenie ma znalezienie pierwszego podziału. Można tego dokonać poprzez przeszukanie wszystkich możliwych podziałów zbioru (na dwie części), ze względu na każdą cechę. Wybieramy ten podział, który daje dwa najbardziej różniące się między sobą podzbiory, albo inaczej – najbardziej zmniejsza zróżnicowanie zbioru badanego. Stosuje się różne miary tzw. *Zanieczyszczenia*. Najpopularniejszy jest wskaźnik Giniego, który jest sumą iloczynów proporcji klas w grupach. Wykorzystywany jest on między innymi (obok innych miar) w metodologii CART *(Classification and Regression Trees).*

Podczas budowy drzewa należy w pewnym momencie podjąć decyzję o zakończeniu dalszego dzielenia (i tworzeniu dalszych węzłów). Teoretycznie drzewo można budować do momentu, aż wszystkie obserwacje zostaną prawidłowo zakwalifikowane (lub zaprognozowane). Takie drzewo jest jednak zazwyczaj bardzo rozległe, a jego liście mogą zawierać nawet pojedyncze obserwacje. Przy budowie drzewa można więc zadać kryterium zatrzymywania rozwoju drzewa, określające minimalna liczbę obiektów w liściu, lub maksymalne dopuszczalne *zanieczyszczenie* klasy, mierzone frakcją „obcych" obiektów.

Zbudowane drzewo zazwyczaj nie jest od razu „optymalnej wielkości". Stosuje się procedurę przycinania. Jeżeli przykładowo drzewo składające się z dziesięciu węzłów prawidłowo klasyfikuje 92% obiektów, a wyeliminowanie jednego węzła zmniejsza trafność wnioskowania do 90%, to najczęściej warto to drzewo zmniejszyć. W tym momencie pojawia nam się problem *kosztów*, obecny w całym procesie budowy i wykorzystania drzew. Koszty błędnego zakwalifikowania obiektu powinny być brane pod uwagę w procesie doboru zmiennych diagnostycznych, ustalania zasad podziału, zatrzymywania budowy drzewa, a także oczywiście jego przycinania. Niekiedy występuje jeszcze problem wyboru drzewa ostatecznego spośród kilku o bardzo podobnych własnościach (np. podobnym błędzie niewłaściwej klasyfikacji). Wówczas zazwyczaj wybiera się drzewo najmniej skomplikowane, czyli posiadające najmniej węzłów. Przy budowie drzewa można wykorzystać dodatkowe informacje o badanym zjawisku. Jeżeli wiemy, że w populacji klasa A występuje dwa razy częściej niż klasa B, to przy losowym „napływie" obiektów do klasyfikacji drzewo powinno identyfikować je w proporcji zbliżonej do 2:1.

Metody budowy drzew klasyfikacyjnych umożliwiają wykorzystywanie cech jakościowych oraz ilościowych skokowych i ciągłych. Metody te poszukują optymalnego podziału zakresu zmienności cechy w ramach każdego węzła.

Tak jak w wielu zagadnieniach Data Mining, zalecane jest, aby badany zbiór obiektów podzielić na dwie części – zbiór uczący oraz zbiór testowy. Drzewo zbudowane na podstawie informacji „wyciągniętych" ze zbioru uczącego, sprawdzamy następnie na zbiorze testowym (jest to przykład tzw. *oceny krzyżowej*).

Sieci neuronowe

Sieci neuronowe to szeroko rozwinięte metody o „ideologii" doskonale pasującej do poszukiwania prawidłowości w masie danych, czyli do data mining. Mamy zmienne *wejściowe* (zwane w innych zagadnieniach *niezależnymi*) oraz zmienne *wyjściowe* (*zależne, prognozowane*). Tworzą one odpowiednio *warstwę wejściową* oraz *warstwę wyjściową* sieci. To, co się dzieje „pomiędzy", tworzy tzw. *warstwę ukrytą*. Warstwa ta składa się z neuronów, które przetwarzają informacje napływające z warstwy wejściowej lub poprzedniej warstwy ukrytej i przekazują do dalszej warstwy ukrytej lub do warstwy wyjściowej. Te wszystkie elementy tworzą architekturę sieci. W każdym neuronie „działa" funkcja aktywacji, która obejmuje *funkcję kombinującą*, która „scala" informacje wejściowe napływające do neuronu, oraz funkcję, która wylicza wartość wyjściową neuronu na podstawie wartości funkcji kombinującej. Ta ostatnia jest zazwyczaj kombinacją liniową wejść. Z kolei funkcja transferująca wartość tej liniowej kombinacji w sygnał wyjściowy ma najczęściej kształt sigmoidalny. Funkcja ta przyjmuje wartości z przedziału [-1,+1], a w swej środkowej części jest zbliżona do funkcji liniowej.

Uczenie sieci neuronowej polega na znajdowaniu „właściwych" wag (współczynników) wykorzystywanych przy integrowaniu informacji wejściowych, tak aby najlepiej przewidywać wartości zmiennej wyjściowej. Stosuje się tu podejście zwane *propagacją wsteczną*. Przy pomocy aktualnych wag wylicza się wartość zmiennej wyjściowej, porównuje się ją z wartością rzeczywistą, oceniając błąd. Następnie zmienia się wagi tak, aby zmniejszyć błąd. Uczenie przerywa się, gdy błędu nie można już istotnie zmniejszyć. Należy pamiętać, że sieć neuronowa może dostarczyć predykcji lub zidentyfikować obiekt, ale nie wyjaśnia mechanizmu, jaki rządzi powiązaniami pomiędzy zmiennymi wejściowymi a wynikowymi.

Analiza skupień

W analizie skupień poszukujemy grup obiektów podobnych, przy czym *a priori* ta liczba grup nie jest znana. Grupowanie powinno być poprzedzone badaniem jednorodności. Jeżeli zbiór obiektów może być uznany za próbę wygenerowaną przez jeden, jednomodalny, wielowymiarowy rozkład prawdopodobieństwa, to oznacza to, że nie ma podstaw do odrzucenia hipotezy o jednorodności, a co za tym idzie nie ma podstaw do dzielenia analizowanego zbioru na podzbiory.

Jeżeli odrzucamy hipotezę o jednorodności, to metody grupowania powinny podzielić badany zbiór na podzbiory, w których jednorodność jest zachowana. W końcowym efekcie dla każdej pary obiektów z analizowanego zbioru możemy stwierdzić, czy obiekty są „podobne" (należące do tego samego podzbioru) czy „niepodobne" (należące do różnych podzbiorów). Niepodobieństwo może oczywiście wynikać z tego, iż jeden z obiektów danej pary jest „lepszy" niż drugi, ale może ono również wynikać stąd, że jest on „inny" (ze względu na strukturę wartości cech).

W zagadnieniach grupowania mamy do czynienia z problemem wyboru cech, ustaleniem systemu wag, normalizacją zmiennych i wyborem miary odległości.

Zasadniczy wybór dotyczy strategii i metody grupowania. Istnieją dwie strategie: hierarchiczna i niehierarchiczna. Ta druga wymaga (w większości przypadków) podjęcia decyzji arbitralnej, dotyczącej bądź liczby podgrup (np. w metodzie k-średnich), bądź tzw. odległości krytycznej, od której zaczyna się „niepodobieństwo".

Hierarchiczne metody aglomeracyjne pozwalają na określenie tzw. hierarchii drzewkowej elementów analizowanego zbioru obiektów. Drzewko połączeń - czyli *dendrogram* - otrzymuje się poprzez krokową aglomerację (łączenie w podzbiory) operacyjnych jednostek taksonomicznych. Na wstępie procedury przyjmuje się, że każdy obiekt stanowi osobną podgrupę. W macierzy odległości poszukuje się najmniejszego elementu spośród leżących poza główną przekątną. Wskazuje on, które podgrupy należy łączyć w kolejnym kroku aglomeracji. Ta najmniejsza odległość określana jest mianem *odległości aglomeracyjnej*. Odległość taka jest „minimalna" w sensie lokalnym, gdyż zazwyczaj jest ona inna przy każdym kolejnym połączeniu. Po połączeniu dwóch wskazanych podgrup należy skorygować macierz odległości. Ubywa z niej jeden wiersz i jedna kolumna - te, których numer odpowiada większemu indeksowi z dwóch łączonych podgrup. Można to uczynić na wiele sposobów i stąd różne wersje hierarchicznych metod aglomeracyjnych. Powszechnie potwierdzona jest najlepsza efektywność metody Warda. W metodzie tej brana jest pod uwagę zmienność wewnątrzgrupowa. Odległość między grupami jest definiowana jako moduł różnicy między sumami kwadratów odległości punktów od środków grup, do których te punkty należą. Metoda ta ma skłonność do tworzenia równomiernego drzewka połączeń, a przy jego podziale powstają grupy o podobnej liczbie obiektów, bez podgrup zawierających pojedyncze obiekty izolowane.

W literaturze opisano około trzydziestu kryteriów zatrzymywania procesu aglomeracji. Wydaje się, że przy decyzji o podziale dendrogramu, czyli przerywaniu procesu aglomeracji i wyborze klasyfikacji ostatecznej, niezbędne jest pewne doświadczenie metodologiczne w sytuacji, gdy nie da się sformułować jednoznacznego kryterium formalnego.

Metody niehierarchiczne wymagają na ogół podejmowania pewnych subiektywnych decyzji, które decydują o klasyfikacji wynikowej. W grupie metod obszarowych przestrzeń klasyfikacji jest dzielona na rozłączne podobszary o kształtach właściwych dla danej metody. Wielkość takiego obszaru jest zazwyczaj ustalana właśnie arbitralnie.

W metodach obszarowych, podobnie jak w strategiach hierarchicznych, decyzja o połączeniu pary obiektów podjęta w trakcie procesu delimitacji nie może być w jego toku zmieniona. Obiekty raz zakwalifikowane razem pozostają wspólnie w tej samej podgrupie do końca „pracy" algorytmu. Inaczej dzieje się w metodach iteracyjnych. Najpopularniejszą metodą z tej grupy jest metoda k-średnich. Niestety każdy algorytm k-średnich wymaga na wstępie określenia liczby podgrup, na które będziemy dzielić nasz zbiór operacyjnych jednostek taksonomicznych. Liczba tych skupisk oznaczana jest literą k i stąd nazwa metody. Po przyjęciu pewnej liczby k wybieramy właśnie k początkowych środków skupisk. Wszystkie obiekty przyporządkowujemy do tych skupisk. Sprawdzamy wartość

funkcji kryterium oceniającej „dobroć" klasyfikacji, (jeżeli wersja metody to przewiduje), a następnie przesuwamy obiekt (lub obiekty) tak, aby poprawić wartość funkcji kryterium.

W literaturze zaproponowano wiele wersji metody k-średnich. Ta różnorodność wynika z przyjmowania różnych:

- sposobów ustalania początkowej konfiguracji środków podgrup,
- kryteriów oceny dobroci klasyfikacji,
- reguł przesuwania obiektów do innych podgrup,
- reguł zatrzymywania procesu poprawiania klasyfikacji.

Prosta i intuicyjnie zrozumiała wersja metody k-średnich przewiduje losowy wybór *k* obiektów z analizowanego zbioru i uznanie ich za wstępne środki analizowanych podgrup. Następnie resztę obiektów przyporządkowujemy do najbliższych środków. Dla każdej tak utworzonej podgrupy wyznaczamy nowy środek, którego współrzędnymi są średnie arytmetyczne ze współrzędnych obiektów należących do tej podgrupy. Z kolei sprawdzamy, czy każdy punkt jest bliżej środka własnej podgrupy, czy też cudzej. W tym drugim przypadku obiekt jest przesuwany do tej grupy, do której środka ma najbliżej. Po przesunięciu obiektów wyznaczamy nowe środki podgrup i całą procedurę powtarzamy, aż do momentu, gdy już żaden obiekt nie da się przesunąć do innej podgrupy (bo dla wszystkich najbliższym środkiem jest środek ich - a nie „cudzej" - grupy).

W każdej wersji metody k-średnich liczba podgrup pozostaje niezmienna, a tylko w toku kolejnych iteracji zmienia się skład tych podgrup. Niedogodność wynikającą z konieczności zdefiniowania liczby podgrup od razu na wstępie można ominąć poprzez przyjmowanie kolejnych, różnych wartości *k* i ocenianie otrzymanych podziałów wynikowych przy użyciu jakiejś miary jakości podziału lub z wykorzystaniem kryteriów merytorycznych.

Z punktu widzenia obliczeń, metodę k-średnich można w pewnym sensie traktować jako "odwrotność" analizy wariancji. Rozpoczynamy od *k* (losowych) skupień, a następnie przenosimy obiekty między tymi skupieniami, mając na celu minimalizację zmienności wewnątrz skupień i maksymalizację zmienności między skupieniami. Jest to analogiczne do "odwrotności" analizy wariancji w tym sensie, że test istotności w analizie wariancji szacuje zmienność międzygrupową w stosunku do zmienności wewnątrzgrupowej, jeśli liczymy test istotności dla hipotezy, że średnie w grupach różnią się między sobą. W grupowaniu metodą k-średnich staramy się przenosić obiekty do i z grup (skupień), aby otrzymać najbardziej istotne wyniki analizy wariancji.

Wyniki analizy grupowania metodą k-średnich (lub każdą inną dającą klasyfikację zbioru obiektów na rozłączne podgrupy) możemy zweryfikować, opisać i zinterpretować poprzez porównywanie średnich wartości cech w podgrupach. Możemy ocenić, na ile nasze skupienia są od siebie różne, i które cechy decydują o tym, że są różne. Analiza wariancji może być tu pomocna przy ocenie (*post-hoc*) dyskryminacyjnej zdolności cech, które wykorzystaliśmy przy grupowaniu.

Modele regresji

Analiza korelacji polega na wyliczaniu miar współzależności zmiennych, testowaniu ich istotności i interpretacji wyników. W takim podejściu korelacje się identyfikuje i mierzy. Stwierdzamy, czy współzmienność ma charakter systematyczny, mierzymy jej siłę i kierunek. Opisywaniem zależności zajmuje się natomiast *analiza regresji*. Tworzymy tu modele, które ilościowo opisują związki między zmiennymi, co pozwala na analizę struktury zależności, znaczenia czynnika losowego oraz umożliwia prognozowanie. Równanie, które opisuje związek między zmiennymi z uwzględnieniem obecności składnika losowego, nazywa się *równaniem* lub *modelem regresji*. W zagadnieniach regresji mamy zdefiniowana tę szczególna zmienną, której wartość trzeba „odgadnąć", wykorzystując wartości zmiennych objaśniających.

W zagadnieniach regresji zazwyczaj należy dobrać właściwy zestaw zmiennych objaśniających oraz przyjąć odpowiednią funkcję łączącą zmienne objaśniające ze zmienna prognozowaną. Podejście data mining zakłada, że dysponujemy dużą liczbą potencjalnych zmiennych objaśniających i nie ma teorii zjawiska, która zdecydowanie determinuje, które zmienne powinny znaleźć się w modelu ostatecznym. Istnieją procedury (np. *regresja krokowa*), które umożliwiają niemal automatyczny dobór zmiennych do modelu. Podobnie niezdeterminowana z góry jest postać analityczna modelu. Na ogół wszyscy zaczynają od modelu liniowego, a następnie poszukują lepszych wersji nieliniowych. Inna droga uwzględniania nieliniowości to modele segmentowe lub lokalne.

Model regresji liniowej w populacji ma postać

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \xi$$

β_0 to wyraz wolny regresji. Zazwyczaj nie podlega on interpretacji. Współczynniki $\beta_1, \beta_2, ..., \beta_k$ to *współczynniki regresji cząstkowej* i ich interpretacja jest niesłychanie ważna dla zrozumienia mechanizmu zjawiska opisywanego przez równanie regresji. Taki współczynnik informuje, o ile przeciętnie zmieni się poziom zmiennej objaśnianej, jeżeli wartość zmiennej objaśniającej, przy której stoi ten współczynnik, wzrośnie o jednostkę, natomiast wartości pozostałych zmiennych objaśniających nie ulegają zmianie. Współczynniki β_j to tak zwane *parametry strukturalne modelu regresji*. O składniku losowym oznaczonym przez ξ *(ksi)* zakładamy, że jest zmienną losową podlegającą rozkładowi normalnemu o wartości przeciętnej 0 (zero) i stałej wariancji równej σ_ξ^2. Graficznym obrazem równania regresji wielu zmiennych jest *hiperpłaszczyzna* regresji.

Jeżeli model opisuje liniową zależność pomiędzy tylko jedną zmienną objaśniającą a jedną zmienną objaśnianą, wtedy jego graficznym obrazem jest *prosta regresji.*

Widzimy, że metody regresji pozwalają zrealizować dwa podstawowe zadania data mining: znalezienie i opisanie prawidłowości występujących w zbiorze danych oraz prognozowanie wartości zmiennej objaśnianej. Niektóre modele regresji umożliwiają też dokonywanie klasyfikacji obiektów.

Metody data mining to dynamicznie rozwijająca się dziedzina. Nie tylko pojawiają się nowe metody, ale również zmieniają się oczekiwania wobec data mining, a także przywraca się znaczenie klasycznych metod statystycznych. Wielu autorów skłania się do twierdzenia, że nierealne jest stworzenie metod całkowicie „automatycznych", wyszukujących prawidłowości, związki, anomalie bez udziału człowieka. Doświadczony badacz może trafnie ukierunkować analizę, ale do efektywnego *miningu* konieczne są metody, które radzą sobie z gigantycznymi plikami danych, wykorzystując zarówno procedury znane od dawna wśród statystyków, jak i nowe pomysły możliwe do zrealizowania dzięki burzliwemu rozwojowi możliwości komputerów.

Literatura

1. Weiss S.M., Indurkhya N., *Predictive Data Mining. A Practical Guide*, Morgan Kaufmann Publishers, San Francisco, 1998.
2. Berry M.J.A., Linoff G.S., *Mastering Data Mining*, John Wiley & Sons, New York, 2000.
3. Hastie T., Tibshirani R., Friedman J., *The elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer, New York – Berlin – Heidelberg, 2001.

