Wprowadzenie do drzew klasyfikacyjnych i regresyjnych

Tomasz Demski

Drzewa klasyfikacyjne w przewidywaniu migracji klientów (churn)

Wraz z rozwojem technologii informatycznej pojawiły się nowe możliwości w zakresie ilości gromadzonych danych oraz szybkości ich przetwarzania. W tej chwili generowanie, przechowywanie czy przesyłanie informacji nie stanowi problemu technicznego. Kluczowym zagadnieniem jest natomiast wydobycie z danych użytecznych informacji, które pomogą podjąć decyzję szybciej i trafniej niż konkurencja.

W odpowiedzi na takie wymagania powstały narzędzia data mining. Drzewa klasyfikacyjne i regresyjne są jedną z najpopularniejszych i najbardziej skutecznych metod data mining, która bardzo często stosowana jest w zapobieganiu migracji klientów (*churn*).

W drzewach klasyfikacyjnych i regresyjnych poszukujemy takich części (segmentów) przestrzeni cech parametrów, w których zmienna zależna przyjmuje tylko pewną określoną wartość z niewielkim błędem. Dla ilustracji rozważmy wykres przedstawiony na Rysunku 1. Widzimy na nim częstość występowania klasy *Tak* w zależności od wartości, jakie przyjmują dwie cechy.

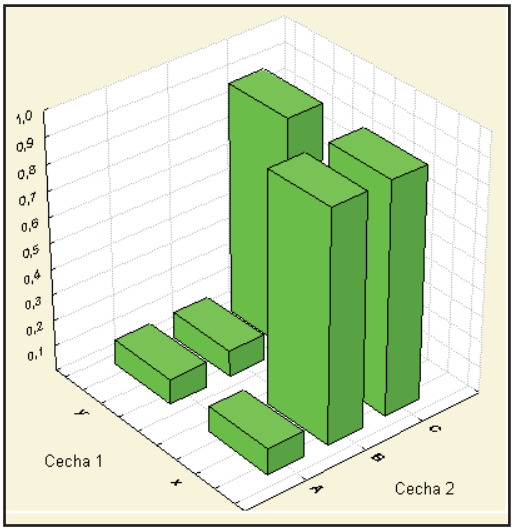
Patrząc na wykres, możemy sformułować następującą regułę:

Obiekt należy do klasy *Tak*:

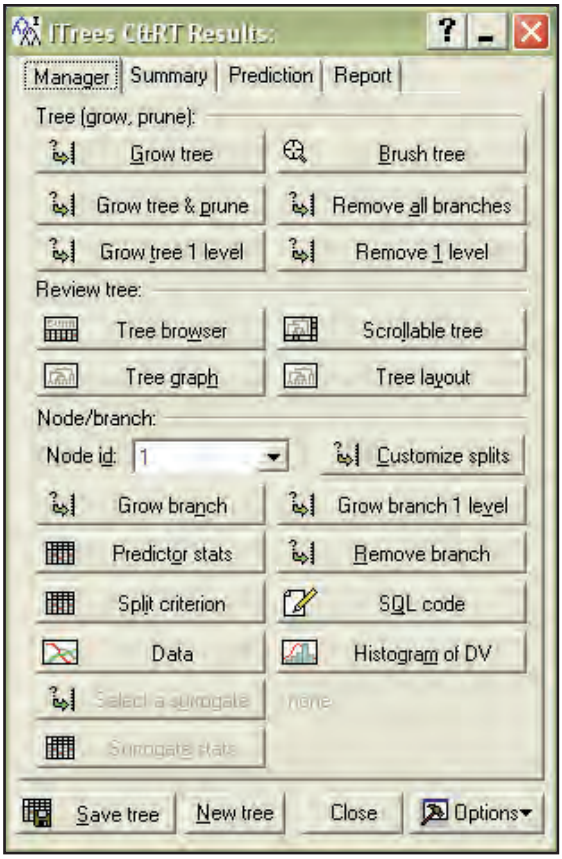
- jeśli *Cecha 2* należy do klasy *C*,
- albo jeśli *Cecha 1* należy do klasy *X* i *Cecha 2* należy do klasy *B*.

Drzewa klasyfikacyjne i regresyjne poszukują podobnych reguł, z tym że są w stanie znaleźć je w bardzo skomplikowanych wielowymiarowych przypadkach, w których analiza „na oko" nie ma szans powodzenia. Drzewa klasyfikacyjne i regresyjne poszukują optymalnego podziału na segmenty, wykonując następujące działania:

1. sprawdzenie dla aktualnie badanego zbioru, czy jest on jednorodny lub czy spełniona jest inna reguła stopu,
2. zbadanie wszystkich możliwych podziałów na rozłączne części,



Rysunek 1. Liczność klas dla różnych wartości predyktorów



Rysunek 2. Panel budowy drzewa klasyfikacyjnego

3. wykonanie najlepszego podziału zbioru na rozłączne części,
4. powtórzenie dla wszystkich zbiorów uzyskanych poprzez wykonanie powyższych czynności.

Jako reguły stopu stosuje się m.in.: minimalną liczność węzła podlegającego podziałom, minimalną liczność węzła powstającego w wyniku podziałów i maksymalną liczbę poziomów drzewa.

Po zakończeniu podziałów wykonuje się jeszcze operację doboru właściwej wielkości drzewa, np. przycinanie (*pruning*). Przycinanie polega na usuwaniu gałęzi drzewa, co wykonujemy automatycznie lub ręcznie, w oparciu o posiadaną wiedzę o celach analizy, jakości pomiaru poszczególnych cech, ograniczeniach stosowania modelu itp. (jest to wiedza, której nie ma w danych i siłą rzeczy analiza danych nie może jej wydobyć).

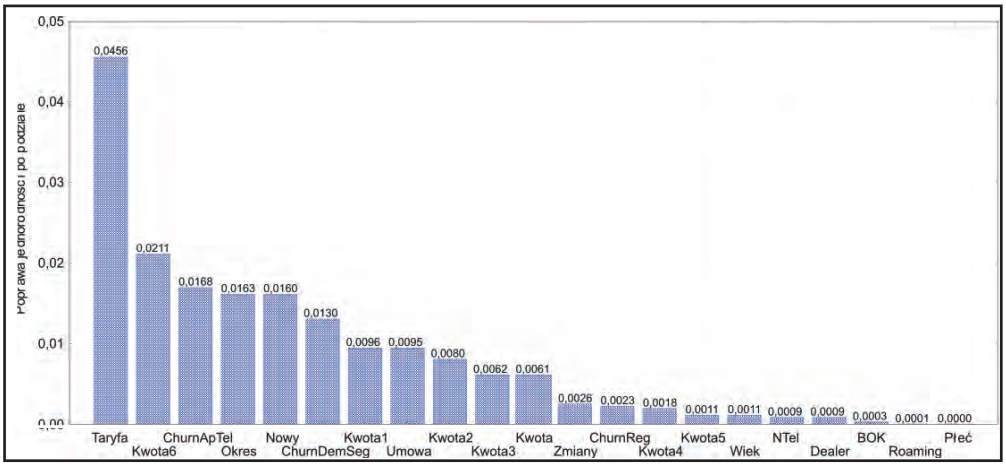
Uzyskane reguły standardowo prezentuje się w postaci drzewa, dzięki czemu są one stosunkowo przejrzyste, nawet gdy drzewo jest spore. Łatwość zrozumienia wyników i możliwość odkrycia łatwych w interpretacji reguł jest jedną z zalet drzew klasyfikacyjnych.

Inne zalety drzew klasyfikacyjnych i regresyjnych to:

- prostota algorytmu, ułatwiająca stosowanie go ze zrozumieniem nawet osobom bez dużego doświadczenia w analizie danych,
- szybkość działania,
- odporność na nietypowe wartości predyktorów,
- odporność na nawet dużą liczbę predyktorów faktycznie niewpływających na badaną zmienną,
- możliwość wychwycenia zależności nieliniowych i interakcji.

Wszystkie te zalety powodują, że dobrze jest zacząć analizę właśnie od tej metody. Z drugiej strony drzewa nie są w stanie opisać tak złożonych zależności jak inne, bardziej skomplikowane matematycznie procedury, np. sieci neuronowe, metoda wektorów nośnych (*Support Vector Machines*) czy drzewa ze wzmacnianiem (*boosted trees*); przegląd metod data mining można znaleźć na stronie www.statsoft.pl/dataminer2.html i w artykule A. Sokołowskiego „Metody stosowane w Data Mining" dostępnym w Czytelni StatSoft Polska (http://www.statsoft.pl/czytelnia/dm/wstepdm.html).

Bardziej szczegółowe wprowadzenie do drzew klasyfikacyjnych i regresyjnych znajduje się w [1].



Rysunek 3. Jakość podziałów wg różnych zmiennych

Migracja klientów a analiza danych

Migracja klientów (*churn*) jest bardzo ważnym problem, zwłaszcza tam gdzie z jednej strony występuje wysoki koszt pozyskania nowych klientów, a z drugiej strony klient bardzo łatwo może odejść do konkurencji, ponosząc przy tym niewielkie lub żadne koszty. W sytuacji gdy rynek nasyca się produktem lub usługą i jest coraz mniej

potencjalnych klientów, znaczenie utrzymania klientów staje się coraz większe.

Metody analizy danych, a w szczególności data mining, stosowane są w zapobieganiu migracji klientów w rozmaity sposób. W szczególności techniki statystyczne wykorzystuje się do badania i monitorowania satysfakcji klienta, wykrywania przyczyn decydujących o zadowoleniu klienta i jego lojalności, monitorowania zmian stopnia satysfakcji klientów. Przegląd i przykłady zastosowań tego typu przedstawiono w pracy [1].

Inna grupa zastosowań analizy danych w kontekście migracji klientów to metody odkrywania wzorców zachowań klientów, pozwalające dostosować ofertę do ich potrzeb.

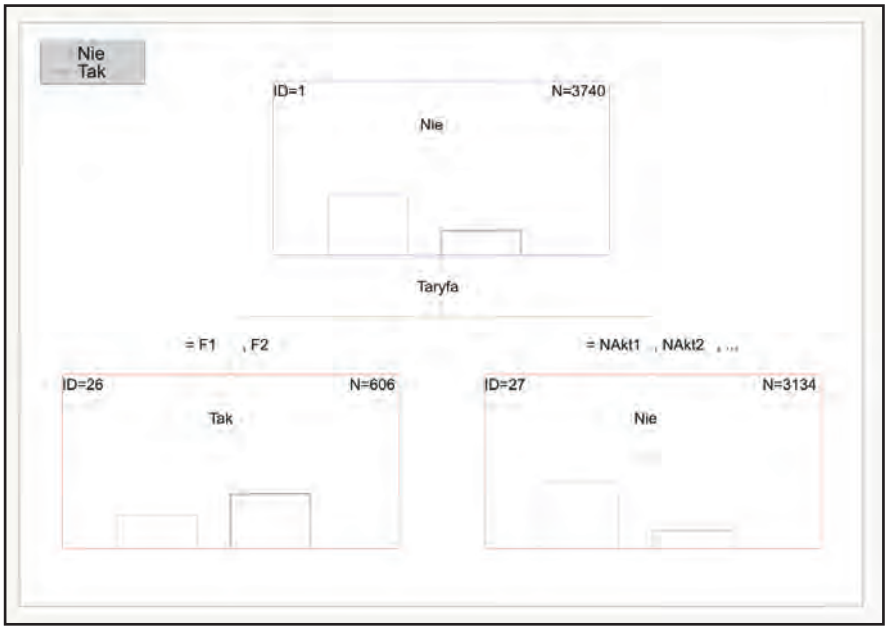
W tym przykładzie zajmiemy się innym (prawdopodobnie najpowszechniejszym) zastosowaniem. Naszym celem będzie znalezienie tych klientów, dla których prawdopodobieństwo odejścia jest największe. Wiedzę taką możemy wykorzystać na dwa sposoby:

1. do modyfikacji oferty, sposobu działania,
2. do podjęcia w stosunku do zagrożonych klientów działań zapobiegających odejściu (np. telefon lub list z ofertą dodatkowych bezpłatnych minut, SMSów itp.).

W drugim przypadku musimy uzyskać prognozę na tyle wcześniej, abyśmy mogli podjąć działania zapobiegawcze, jeszcze **przed** podjęciem przez klienta ostatecznej decyzji. Wpływa to oczywiście na dane wykorzystywane w analizie – do modelowania wykorzystujemy dane poprzedzające odejście klienta na przykład o 1 miesiąc.

Przykład

W naszym przykładzie zajmiemy się sytuacją, gdy dane zostały już zgromadzone i sprawdzone. Innymi słowy, zajmiemy się „czystą" analizą danych. Do przewidywania odejść zastosujemy moduł *Drzewa interakcyjne* (*Interactive trees*) systemu *STATISTICA* Data Miner, ponieważ umożliwiają one zrozumienie związków między zmiennymi i dają pełną kontrolę nad budową drzewa.

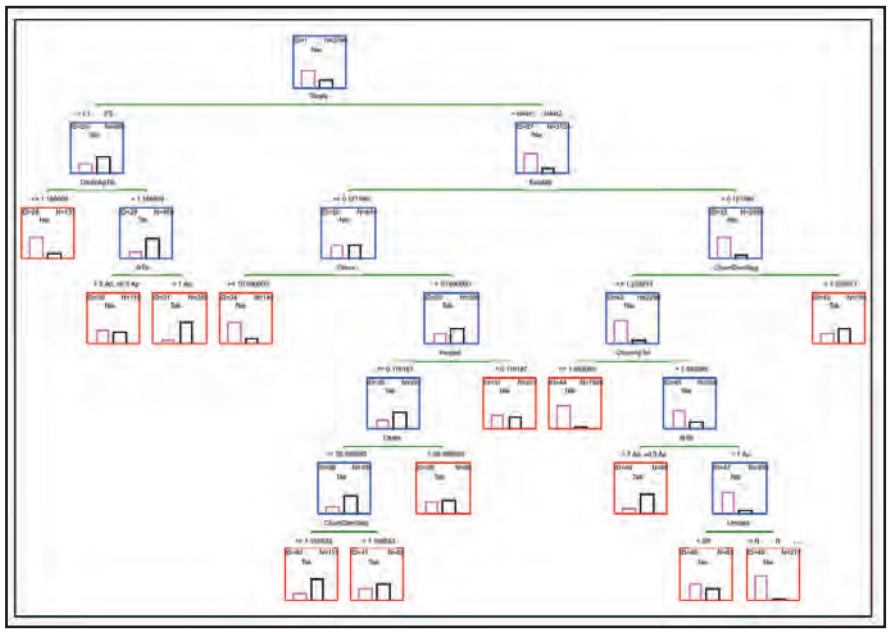


Rysunek 4. Pierwszy podział

Wróćmy teraz do naszego zadania. Problem dotyczy przewidywania migracji pewnego typu klientów operatora telefonii komórkowej.

W pliku danych znajdują się informacje o:

- cechach osobistych i demograficznych (np. wiek, region zamieszkania),
- rodzaju umowy, taryfy,
- interakcjach z dostawcą (np. czy ostatnio był kontakt z Biurem Obsługi Klienta),



Rysunek 5. Drzewo klasyfikacyjne

- wysokości rachunków w ciągu ostatnich 6 miesięcy oraz proporcja tej kwoty przypadająca na poszczególne miesiące,
- częstości odejść dla klientów posiadających podobny aparat telefoniczny, podobne cechy demograficzne itp.

Informacje te wykorzystamy jako *predyktory* w modelu przewidującym, dla których klientów występuje największe zagrożenie odejściem.

Zauważmy, że w praktyce często stosuje się wskaźniki obliczone jako skomplikowane przekształcenie oryginalnych danych. W naszym przykładzie użyjemy zmiennych, które mają jasną interpretację.

Ponieważ frakcja odejść w całej zbiorowości wynosiła 1,2% przed wykonaniem analizy, konieczne było wylosowanie próby o bardziej zrównoważonych proporcjach. W efekcie do modelowania wykorzystamy 4490 przypadków, z których w około 30% wystąpiła *migracja klientów*.

Do budowy drzewa wykorzystamy algorytm C&RT. Jako najmniejszą liczbę obiektów w węźle uzyskanym w wyniku podziału ustawiamy 60, a dopuszczalną liczbę poziomów drzewa ograniczamy do 25.

Dane, którymi dysponujemy, losowo podzielimy na dwie próby: uczącą i testową. Próba ucząca zostanie wykorzystana do zbudowania drzewa, a testowa posłuży nam do sprawdzenia, czy model nie nauczył się danych „na pamięć" i czy będzie się nadawał do przewidywania odejść dla nowych danych. Podział danych został przeprowadzony losowo, a do próby uczącej trafiło około 75% wszystkich przypadków.

Po określeniu parametrów analizy rozpoczynamy budowę drzewa. Na Rysunku 2 widzimy panel sterowania budową drzew – oczywiście możemy zbudować od razu drzewo automatycznie. My jednak przyjrzymy się podziałom, tak aby wybrać najlepsze.

Zacznijmy od pierwszego podziału. Na Rysunku 3 widzimy wskaźniki jakości dla podziałów według wartości wszystkich predyktorów. Dokładniej rzecz biorąc, jest to informacja, o ile zbiory powstałe w wyniku podziału są bardziej jednorodne od wejściowego zbioru.

Z wykresu wynika, że najlepszy podział będzie wykorzystywał zmienną *Taryfa*. W przypadku gdy różnice między poprawą jednorodności nie byłyby tak wyraźne, moglibyśmy wybrać inną zmienną, niż ta wskazana przez analizę danych, wykorzystując „zewnętrzną" wiedzę. Przykładowo, jedna ze zmiennych może dawać minimalnie gorszy podział, ale być dokładniej mierzona lub łatwiej i szybciej uzyskiwana. Ponieważ znaleziony przez program podział jest odpowiedni, akceptujemy go. Na Rysunku 4 widzimy drzewo klasyfikacyjne uzyskane w wyniku podziału.

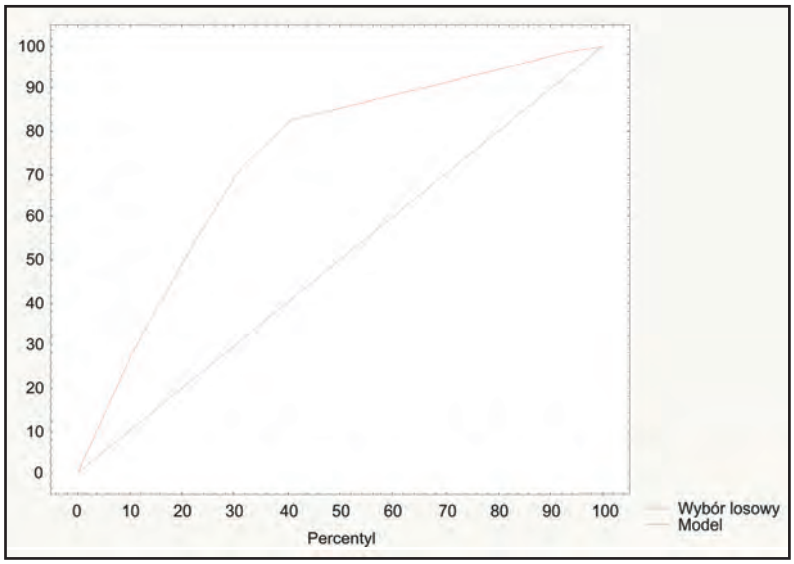
W podobny sposób możemy śledzić i kontrolować cały proces budowy drzewa. Kompletne drzewo widzimy na Rysunku 5. Jest ono dosyć duże: składa się z 12 węzłów wewnętrznych i 13 końcowych (liści).

Zwróćmy uwagę na węzeł numer 31. Wśród klientów zakwalifikowanych do tej grupy, aż w 302 przypadkach na 350 doszło do odejścia. Jednocześnie są to „dziwni" klienci: korzystają z taryfy F1 lub F2, która jest pomyślana dla osób korzystających z wielu aparatów telefonicznych, a jednocześnie używają jednego telefonu. Zauważmy, że używane przez nich aparaty były dosyć stare i wśród ich użytkowników było sporo odejść (zmienna *ChurnApTel* >1,1669).

Grupę tę możemy zinterpretować jako sprytnych klientów: podpisujących umowę na dwa aparaty, dostających na promocyjnych warunkach nowy telefon i natychmiast rezygnujących z jednego z abonamentów. Tę wiedzę możemy wykorzystać np. do „uszczelnienia" umów na taryfy F1 i F2, aby uniemożliwić takie ich wykorzystanie.

Do oceny jakości modelu wykorzystamy stopę błędnych klasyfikacji w próbie uczącej i testowej: wynoszą one 17,4% dla zbioru uczącego i 19,0% dla zbioru testowego, co wskazuje, że model nie jest przeuczony i będzie poprawnie radził sobie z nowymi danymi.

Bardzo użytecznym narzędziem oceny i udoskonalenia modelu jest tzw. wy-



Rysunek 6. Wykres zysku (*gain chart*)

kres zysku (*gains chart*) przedstawiony na Rysunku 6.

Wykres ten pokazuje nam, jaki procent wszystkich odejść przewidzimy, jeśli wybierzemy te obserwacje, dla których prawdopodobieństwo odejścia wynikające z modelu jest większe od kolejnych „górnych" percentyli rozkładu prawdopodobieństwa przewidywanego przez model (zazwyczaj 10., 20., 30. itd.). W tym wypadku górny percentyl rzędu n jest rozumiany jako wartość, od której większe jest n% wszystkich obserwacji. Zauważmy, że z powyższego wykresu wynika, że jeśli wybierzemy tylko te obserwacje, dla których prawdopodobieństwo wynikające z modelu jest większe od 40. percentyla, to przewidzimy 80% wszystkich odejść. Jeśli działanie zapobiegające odejściu klienta jest drogie (np. wysłana pocztą oferta na przedłużenie umowy w zamian za duży pakiet bezpłatnych minut) lub kłopotliwe i czasochłonne, to bardzo korzystne będzie zawężenie działania *antychurnowego* właśnie do wybranej za pomocą wykresu zysku grupy klientów (w naszym wypadku tych, dla których prawdopo dobieństwo odejścia jest większe od górnego 40. percentyla).

Jeżeli uznamy, że nasz model jest zadowalający, to zapewne będziemy chcieli go użyć w praktyce. Dosyć często najlepszym rozwiązaniem jest wygenerowanie kodu dla modelu i przeniesienie go do innego systemu – np. bazy danych o klientach. W *STATISTICA* Data Miner możemy zapisać model drzew w postaci kodu C, XML (a właściwie dialektu XML: PMML, który został zaprojektowany do przenoszenia modelu data mining między różnymi systemami), SQL i STATISTICA Visual Basic. ■

Literatura

[1] Analiza satysfakcji i lojalności klientów, 2003, StatSoft Polska
[2] Berry M.J.A., Linoff G., 2000, Mastering data mining, John Willey & Sons
[3] Berson, A., Smith, S., Thearing, K., 1999, Building Data mining Applications for CRM, McGraw-Hill

Tomasz Demski pełni funkcję kierownika Działu Technicznego StatSoft Polska. Kontakt: t.demski@statsoft.pl.