



ANALIZA MIGRACJI KLIENTÓW (*CHURN ANALYSIS*)

Mariusz Łapczyński, Uniwersytet Ekonomiczny w Krakowie

Wprowadzenie do analizy migracji klientów (*churn analysis*)

Termin *churn* jest stosowany w branży telekomunikacyjnej i oznacza odejście klientów do konkurencyjnych operatorów sieci komórkowych. W praktyce może oznaczać utratę klientów na rzecz konkurencji. Istnieją dwa podejścia do szacowania wskaźnika odejścia¹¹. Pierwsze opiera się na metodzie RFM (*recency, frequency, monetary*), która w dużym skrócie polega na zbadaniu, kiedy ostatnio dokonano zakupu (*recency*), z jaką częstotliwością dokonywano zakupów w badanym okresie (*frequency*) i jaka była wartość poszczególnych zakupów (*monetary*). Analizuje się tu wskaźnik *recency*, który jest tożsamy ze wskaźnikiem odejścia.

Drugie podejście to metoda VAL (*value, activity, loyalty*), w której szacuje się indywidualne zachowanie klienta w oparciu o większą liczbę przypadków. O ile w pierwszej metodzie o wysokości wskaźnika odejścia dla pojedynczego klienta decydował wskaźnik *recency* dla tej konkretnej osoby, o tyle w drugiej metodzie wskaźnik odejścia szacowany jest z większej liczby przypadków (np. danego segmentu).

Wyróżnia się trzy rodzaje migracji klientów¹²:

- ◆ dobrowolne migracje;
- ◆ wymuszone migracje;
- ◆ oczekiwane (przewidywalne) migracje.

Dobrowolne migracje klientów oznaczają świadome zerwanie współpracy klienta z przedsiębiorstwem.

¹¹ M.A.P.M. Lejeune, *Measuring the impact of data mining on churn management*, Internet Research: Electronic Networking Applications and Policy, Volume 11, Number 5/2001.

¹² M.J.A. Berry, G.S. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, Wiley Publishing, 2004, s. 118.



Pomimo że analiza migracji klientów jest utożsamiana ze sferą usług (branża telekomunikacyjna, bankowość detaliczna), to coraz częściej mówi się o niej także w kontekście innych obszarów i branż¹³:

- ◆ handlu detalicznego,
- ◆ branży ubezpieczeniowej,
- ◆ produkcji komputerów,
- ◆ branży motoryzacyjnej.

Wymuszone migracje mają miejsce wówczas, gdy koniec współpracy następuje z inicjatywą przedsiębiorstwa. Przyczyną są najczęściej nieuregulowane płatności przez klienta. Z oczekiwaną migracją mamy z kolei do czynienia wtedy, gdy klienci w naturalny sposób przestają być klientami firmy. Dorastające dzieci nie potrzebują już środków pielęgnacyjnych dla niemowląt, konsumenci przechodzą do kolejnych faz cyklu życia rodziny i zmieniają swoje zwyczaje zakupowe itp.

Inna klasyfikacja podejść do analizy migracji klientów jest związana ze sposobem definiowania zmiennej zależnej. Możliwe są dwa warianty analizy. W pierwszym z nich zmienna zależna przyjmuje dwie wartości: 1 = klient pozostanie klientem firmy oraz 0 = klient zrezygnuje z usług firmy. Budowa modelu predykcyjnego z tak zoperacjonalizowaną zmienną zależną wymaga przyjęcia odpowiedniego horyzontu czasowego – zwykle 60-90-dniowego. Okres ten nie może być zbyt krótki, gdyż potrzebny jest czas na podjęcie działań zapobiegawczych, zachęcających klienta do kontynuowania współpracy. Binarny charakter zmiennej zależnej sprawia, że zasadniczym celem badacza jest oszacowanie prawdopodobieństwa przynależności do każdej z klas. Najczęściej wykorzystuje się w tym celu drzewa klasyfikacyjne, sieci neuronowe bądź regresję logistyczną. W modelach predykcyjnych z binarną zmienną zależną zwraca się uwagę przede wszystkim na wskazanie grupy klientów, którzy mają zamiar zrezygnować z usług firmy.

Inaczej przedstawia się sytuacja w drugim podejściu, gdzie celem badacza jest predykcja długości czasu trwania współpracy. Tak zoperacjonalizowana zmienna zależna (czas współpracy) pozwala „wydobyć” z modelu większą ilość informacji. Może ona być przykładowo wykorzystana do szacowania wartości życiowej klienta lub do budowy lojalnościowych modeli scoringowych (im dłuższy okres współpracy, tym większa liczba punktów). Narzędziem wykorzystywanym do budowy modeli predykcyjnych jest analiza przeżycia. Największym utrudnieniem w tym podejściu jest nieprzewidywalność zmian w otoczeniu, które potencjalnie mogą determinować zachowania nabywców.

Migracje klientów w branży telekomunikacyjnej

Klientów uznaje się za nielojalnych (*churnerów*), kiedy zrywają umowę z dotychczasowym dostawcą usług telekomunikacyjnych i zostają klientami konkurencji. W branży

¹³ W. Buckinx, D. Van den Poel, *Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Concractual FMCG Retail Setting*, Working Paper, Ghent University, May 2003, s. 4.



telekomunikacyjnej istnieje podział klientów odchodzących na absolutnych – tych, którzy zakończyli współpracę i korzystają z usług konkurencji, oraz liniowych (usługowych), czyli tych, którzy zrezygnowali tylko z niektórych linii (usług), ale wciąż pozostają klientami firmy.

Szacuje się, że w Stanach Zjednoczonych liczba *churnerów* w grupie klientów korzystających z rozmów międzymiastowych wynosi około 2 miliony rocznie. W Europie wśród klientów operatorów sieci komórkowych odsetek ten wynosi miesięcznie 8-12%. Koszt utraty klienta to dla operatora sieci kwota rzędu 500 euro. Powszechnie wiadomo, że koszt pozyskania nowych klientów znacznie przewyższa koszt utrzymania dotychczasowych, jednak w branży telekomunikacyjnej jest to szczególnie widoczne. Nowi abonenci często odchodzą do konkurencji, zanim firmie zwróci się koszt ich pozyskania. Dlatego coraz częściej mówi się nie tylko o analizie migracji klientów (*churn analysis*), ale o zarządzaniu migracjami (*churn management*). Problem migracji klientów zaczyna również dotyczyć firm polskich, które ze względu na otwarcie rynków muszą rywalizować z niejednokrotnie bardzo silnymi konkurentami z zagranicy.

Wykorzystanie narzędzi *data mining* w zarządzaniu migracjami klientów ma dwa podstawowe cele: predykcję tychże migracji oraz zrozumienie przyczyn nielojalności. Budowa modelu predykcyjnego jest związana z próbą zatrzymania klienta. Przedsiębiorstwa stosują różnorodne bodźce, których celem jest zachęcenie konsumenta do przedłużenia umowy. Poznanie przyczyn odejścia pozwala z kolei na optymalizację oferty w taki sposób, aby wyeliminować te, które w największym stopniu powodują niezadowolenie klientów. Ścisłe powiązana z analizą migracji jest analiza rentowności, która pomaga zidentyfikować najbardziej dochodowych klientów. Przedsiębiorstwo nie ponosi bowiem dużej straty jeśli *churnerami* zostają klienci niskorentowni lub nierentowni.

Zestaw zmiennych niezależnych w analizie migracji w branży telekomunikacyjnej obejmuje zazwyczaj:

- ◆ cechy demograficzne klientów (wiek, płeć, miejsce zamieszkania, stan cywilny);
- ◆ statystyki rozmów (długość rozmów w różnych porach dnia, liczba rozmów międzymiastowych oraz liczba rozmów lokalnych);
- ◆ informacje o rachunku (kwoty wydawane na rozmowy międzymiastowe i lokalne);
- ◆ dodatkowe informacje o usługach (np. dodatkowe postanowienia w umowie, zmodyfikowane taryfy);
- ◆ produkty i usługi nabywane przez klientów (np. połączenia z Internetem, zakup usług multimedialnych itp.);
- ◆ informacje o skargach klienta (dotyczących wysokości rachunków, szybkości połączeń, nieprawidłowości związanych z działaniem określonych usług/produktów).

Badania przeprowadza się zwykle dla każdego segmentu klientów z osobna, przy czym w trakcie analizy napotyka się na szereg trudności. Przykładowo klienci biznesowi świadomie różnicują dostawców usług, zawierając umowy z kilkoma operatorami (dostawcami usług) jednocześnie. Problem ten dotyczy także klientów indywidualnych, którzy mają



możliwość wyboru innej firmy dla różnego rodzaju usług telekomunikacyjnych (np. karta Telegrosik do rozmów międzymiastowych). Utrudnieniem w budowie modeli predykcyjnych są także okresowe zmiany warunków umowy dla abonentów oraz dosyć częste w tej branży przymusowe migracje.

Wydatki europejskich firm z branży telekomunikacyjnej na analizę migracji klientów wzrosły z 0,5 mld dolarów w 1999 r. do 3,5 mld dolarów w 2004 r. Pomimo tak dużych kosztów zauważa się szereg korzyści, jakie można odnieść dzięki zarządzaniu migracjami:

- ◆ maleją nakłady na pozyskiwanie nowych klientów,
- ◆ zwiększa się rentowność dotychczasowych klientów,
- ◆ wzrasta sprzedaż usług dodatkowych dla klientów długo pozostających w sieci,
- ◆ zwiększa się zaufanie inwestorów do przedsiębiorstwa.

Migracje klientów w bankowości detalicznej

Bankowość detaliczna jest obszarem, w którym klienci korzystają z usług jednej firmy przez bardzo długi okres czasu. Wymusza to na bankach stosowanie długoterminowych strategii kontaktów z klientami, co często przejawia się tym, że banki udostępniają bezpłatne konta młodym osobom (uczniom, studentom) w nadziei na zyski, jakie czerpać z nich będą w przyszłości. Owa strategia jest ściśle związana z pojęciem – wartość życiowa klienta (LTV, *life time value*), które oznacza sumę zysków netto, jakie osiąga firma podczas współpracy z jednym klientem przez cały okres trwania tej współpracy. Oszacowanie wartości klienta tuż po tym, jak zmieni bank, jest sprawą stosunkowo łatwą. Znacznie trudniej jest zdefiniować wskaźnik LTV w trakcie trwania współpracy lub – co ważniejsze – przewidzieć, jaka będzie wartość klienta zanim wejdzie on w fazę największej aktywności zawodowej.

Analiza rentowności klienta pozwala nie tylko na oszacowanie obecnych i przyszłych dochodów, ale również na podział klientów na segmenty wg kryterium „zyskowności”. Owa segmentacja może być później wykorzystana w optymalizowaniu oferty, stosowania bodźców lojalnościowych czy zarządzaniu migracjami.

Definicja klienta nielojalnego (*churnera*)¹⁴ w bankowości detalicznej jest nieco inna niż w branży telekomunikacyjnej, gdzie abonent może stosunkowo łatwo i często podpisać umowę z nowym operatorem, nie rezygnując natychmiast z usług dotychczasowego. W bankowości detalicznej, klienci zazwyczaj wiążą się z kilkoma bankami w ciągu całego życia, przy czym współpraca z każdym trwa po kilka lub kilkanaście lat. Nacisk na predykcję *churnerów* jest bardzo duży, gdyż nakłady na pozyskanie klienta zwracają się dopiero po długim okresie czasu. Działy marketingu oczekują zatem od badaczy/analityków informacji o prawdopodobieństwie odejścia klienta do konkurencji. Podobnie jak w innych branżach, umożliwia to podjęcie odpowiednich kroków (różnego rodzaju bodź-

¹⁴ W branży finansowej synonimem terminu *churn analysis* jest termin *attrition analysis*.



ców, zachęt) w celu powstrzymania migracji. Działania te dotyczą oczywiście tylko klientów zaliczanych do tzw. zyskownych segmentów.

Zestaw zmiennych niezależnych w analizie migracji w bankowości detalicznej obejmuje zwykle:

- ◆ wysokość wpłat na konto;
- ◆ wysokość wypłat z konta;
- ◆ informację o usługach, z jakich korzysta klient;
- ◆ informacje o właścicielu konta;
- ◆ wysokość dochodów.

Budowa modeli predykcyjnych wymaga od badacza przewyciężenia kilku trudności. Jedną z nich jest np. identyfikacja klienta odchodzącego. Klient zostaje uznany za *churnera* w momencie, gdy likwiduje konto w banku. Oznacza to, że o wiele łatwiej rozpoznać go *post factum*. Predykcja klientów nielojalnych oparta na wskaźniku LTV jest niestety trudna, zwłaszcza w sytuacji, gdy wartość dokonywanych przez nich transakcji jest niezbyt wysoka.

Drugie utrudnienie jest związane z doбором odpowiednich przypadków. Z analizy należy bowiem wyłączyć m.in.: klientów posiadających kredyt hipoteczny albo klientów, którzy zaciągnęli wysoki kredyt konsumpcyjny. Zdarza się wówczas, że banki wymuszają na nich założenie konta osobistego (klienci stają się „przymusowo lojalni”).

Kolejna trudność ma charakter wyłącznie analityczny i dotyczy tzw. problemu nie zrównoważonych klas. Binarna zmienna zależna (lojalny/nielojalny) często charakteryzuje się bardzo nie zrównoważonym odsetkiem obserwacji przypadającym na poszczególne kategorie (np. 96% - 4%). Problem ten jest szczególnie dotkliwy, jeśli celem badacza jest charakterystyka i predykcja przypadków należących do mniej licznej klasy. Podejścia umożliwiające redukcję bądź eliminowanie tego problemu zostaną omówione w dalszej części artykułu.

Migracje klientów w handlu detalicznym

Analiza migracji klientów w branży telekomunikacyjnej bądź bankowości detalicznej pozwala względnie łatwo ustalić moment, w którym klient przestaje być lojalny. Najczęściej wiąże się to z zerwaniem umowy lub niepodpisaniem nowej.

W handlu detalicznym, a w szczególności w przypadku dóbr szybko rotujących (FMCG) sytuacja jest bardziej skomplikowana. Mamy tu do czynienia z częściowymi *churnerami*, czyli osobami, które pozostając klientami sklepu A, realizują część zakupów w konkurencyjnym sklepie B. Istnieje zagrożenie, że „częściowa” nielojalność zamieni się w długim okresie czasu w całkowitą emigrację klienta.

Podobnie jak w innych branżach, analiza migracji powinna tu zostać zawężona wyłącznie do segmentów najbardziej zyskownych. Identyfikacja tych segmentów opiera się na dwóch



kryteriach: częstotliwości zakupów i czasie, jaki upłynął między zakupami (*interpurchase rate*, w skrócie IPT). Wartość pierwszego wskaźnika przekraczająca średnią arytmetyczną dla wszystkich przypadków informuje o lojalności klientów. Równe wartości drugiego wskaźnika informują z kolei o regularności zakupów w danym sklepie i również przesądzą o zakwalifikowaniu klientów do grupy lojalnych. Warto podkreślić, że obie miary nie informują o wartości koszyka zakupów, a jedynie o potencjalnej wartości tej grupy klientów.

O tym, czy klienta uznaje się za częściowo nielojalnego, decydują wartości obu miar – malejąca i niższa od średniej częstotliwość zakupów oraz duża wartość odchylenia standardowego dla wskaźnika IPT. Ta ostatnia świadczy o dużej zmienności, a tym samym o braku regularności dokonywania zakupów w danym sklepie.

Budując modele predykcyjne na potrzeby handlu detalicznego, można wykorzystać następujące zmienne niezależne:

- ◆ czas między kolejnymi zakupami (liczba dni od ostatnich zakupów, średnia liczba dni w badanym okresie, odchylenie standardowe dla tej średniej, iloraz odchylenia standardowego i średniej);
- ◆ częstotliwość zakupów (liczba wizyt w badanym okresie, iloraz liczby wizyt i długości okresu współpracy);
- ◆ wartość koszyka (całkowita kwota wydana w badanym okresie, iloraz tej kwoty i długości okresu współpracy, liczba wizyt, w czasie których kwota wydana na zakupy była wyższa od średniej);
- ◆ kategorię nabywanych produktów (zmienne binarne informujące o fakcie dokonania zakupu produktów z poszczególnych działów: owoce, ryby, mięso, pieczywo, alkohol, napoje, detergenty itp.);
- ◆ markę nabywanych produktów (marka własna, marki krajowe);
- ◆ okres współpracy (liczba dni od dnia pierwszych zakupów);
- ◆ czas dokonywania zakupów (pora dnia, pora dnia podczas ostatniej wizyty);
- ◆ formy płatności (kwoty regulowane za pomocą gotówki, kart kredytowych, kart debetowych, czeków, bonów Sodexo, bonów lojalnościowych);
- ◆ informację o instrumentach promocyjnych (liczba wykorzystanych bonów promocyjnych z danego sklepu, liczba odwiedzin od dnia, kiedy po raz pierwszy skorzystano z bonu promocyjnego, liczba zebranych punktów w programie lojalnościowym itp.);
- ◆ zmienne demograficzne (wielkość gospodarstwa domowego, język używany w domu, kod pocztowy, posiadanie zwierząt, informacja o braku danych demograficznych).

Problem niezrównoważonych klas w budowie modeli *churnowych*

W trakcie budowy modelu predykcyjnego analitycy często mają do czynienia z tzw. problemem niezrównoważonych klas (*imbalanced class problem*), który polega na tym, że



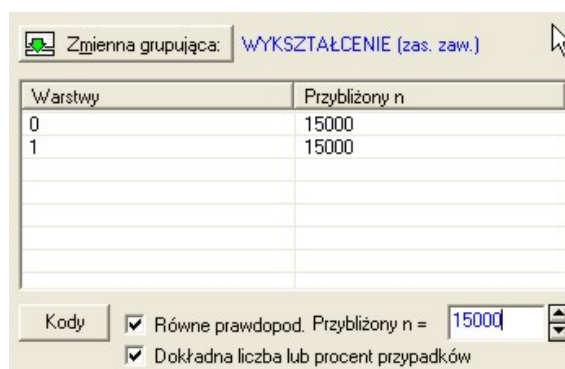
liczba obserwacji należących do jednej z klas jest znacznie mniejsza od liczby obserwacji z drugiej klasy. Klasyczne narzędzia analityczne są zazwyczaj nieskuteczne do tego rodzaju danych, ponieważ ich celem jest budowa modelu minimalizującego ogólny błąd klasyfikacji bez koncentrowania się na którymś (najlepiej tym mniej liczny) z wariantów.

Ogólnie rzecz ujmując, wyróżnia się dwa podejścia analityczne w przypadku zbiorów obserwacji z nierównymi klasami. Jedno z nich polega na wykorzystaniu tzw. wrażliwych algorytmów (*sensitive learning*), które przypisują wysoki koszt błędnej klasyfikacji klasie mniej licznej. Drugie podejście polega na losowej zmianie struktury próby uczącej (*resampling*).

Losowa zmiana struktury próby uczącej

W przypadku zmiany struktury próby uczącej możliwe są następujące strategie:

1. Zmniejszanie liczebności klasy bardziej licznej (*down-sizing, down-sampling*).
2. Zwiększanie liczebności klasy mniej licznej poprzez zwielokrotnienie przypadków z tego podzbioru (*over-sampling, up-sampling*).
3. Podejście łączące strategię pierwszą i drugą.



Rys. 1. Pole ustawień prawdopodobieństw a priori w algorytmie CART.

Pierwszą i drugą strategię nazywa się jednostronnymi technikami próbkowania (*one-sided sampling techniques*), zaś strategię kombinowaną – dwustronną techniką próbkowania (*two-sided sampling technique*). Zmianę struktury próby uczącej można bardzo łatwo przeprowadzić używając modułu „Podzbiór” w menu „Dane” (rys. 1).

Wykorzystanie wrażliwych algorytmów (*sensitive learning*) – algorytm CART

Przed przystąpieniem do analizy za pomocą drzew klasyfikacyjnych CART¹⁵ badacz powinien ustalić prawdopodobieństwa *a priori* pojawienia się klas. Jeżeli prawdopodobieństwa te są szacowane z danych (próby uczącej) tzn., że nacisk na predykcję poszczególnych klas będzie proporcjonalny do występowania tych klas w analizowanym zbiorze obserwacji. Przykładowo: jeśli w bazie danych tylko 5% przypadków to potencjalni klienci, wówczas analiza rozpoczyna się od potraktowania całej bazy jako zbioru potencjalnych nie-klientów

¹⁵ Breiman L. i in., *Classification and Regression Trees*. Chapman and Hall, New York 1984.



(z 5-procentowym błędem). Zaleca się, aby pierwszy model¹⁶ zbudować przy równych prawdopodobieństwach *a priori*, gdyż wtedy każda z klas (każdy wariant zmiennej zależnej) jest traktowana jednakowo.

Jeśli badaczowi zależy na predykcji jednej z klas bardziej niż na predykcji pozostałych, to prawdopodobieństwo *a priori* dla tej właśnie klasy powinno być znacznie wyższe (rys. 2). Sytuacja taka ma często miejsce w badaniach empirycznych, gdzie tylko nieznaczny odsetek baz danych to potencjalni klienci, wiarygodni kredytobiorcy, klienci nielojalni, osoby, które pomyślnie przeszły kurację itp.



Rys. 2. Pole ustawień prawdopodobieństw *a priori* w algorytmie CART.

Podobny skutek można osiągnąć, zmieniając koszty błędnych klasyfikacji (*misclassification costs*), które w większości programów komputerowych są ustawione domyślnie jako równe. Wartości w poniższych tabelach (rys. 3) są względnymi kosztami błędnego sklasyfikowania przypadku z jednej grupy jako przypadku z drugiej grupy. Wartość 5 (w tabeli po prawej stronie) oznacza względną „karę”, jaką ponosi analityk w przypadku, gdy przypadek z klasy B zostanie sklasyfikowany jako przypadek należący do klasy A. Ogólnie rzecz ujmując, zwiększenie kosztów błędnych klasyfikacji dla danej klasy (danego wariantu zmiennej zależnej) zwiększa trafność predykcji tej klasy.

**Koszty błędnych klasyfikacji -
RÓWNE**

Klasa	Klasa 0	Klasa 1
0	1	1
1	1	1

**Koszty błędnych klasyfikacji -
USTALONE**

Klasa	Klasa 0	Klasa 1
0	1	5
1	1	1

Rys. 3. Przykład ustawień kosztów błędnych klasyfikacji.

Wykorzystanie wrażliwych algorytmów (sensitive learning) – algorytm CART – agregacja bootstrapowa modeli predykcyjnych (bagging)

*Bagging*¹⁷ to akronim od słów „bootstrap aggregating”, który w języku polskim funkcjonuje pod nazwą agregacji bootstrapowej¹⁸. Jest to metoda generowania wielu wersji modeli drzew klasyfikacyjnych lub regresyjnych, które zostają ostatecznie zastąpione jednym

¹⁶ W *data mining* zaleca się budowę kilku modeli z wykorzystaniem różnych narzędzi, różnych algorytmów i różnych opcji, a następnie porównanie otrzymanych wyników i wybór najlepszego z nich.

¹⁷ L. Breiman, *Bagging predictors*, Technical Report No. 421, Department of Statistics, University of California, Berkeley, September 1994.

¹⁸ E. Gatnar, *Nieparametryczna metoda dyskryminacji i regresji*, PWN Warszawa 2001, s. 120.



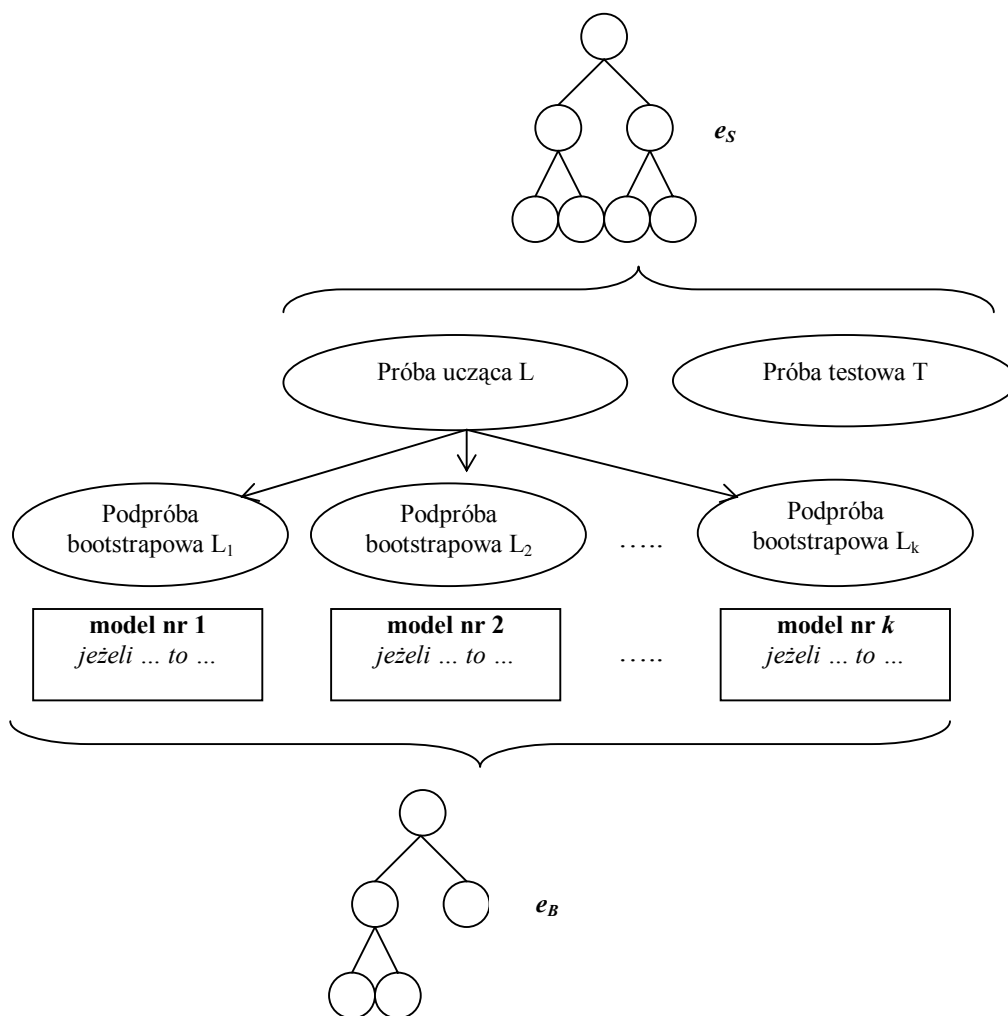
zagregowanym modelem predykcyjnym. Jeśli zmienna zależna jest ilościowa, wówczas zagregowane drzewo powstaje poprzez uśrednienie modeli składowych, jeśli natomiast zmienna ta jest jakościowa, to agregacja oparta jest na częstości występowania poszczególnych klas.

Pojedyncze drzewa są budowane na podstawie danych z podprób bootstrapowych, czyli podzbiorów wylosowanych z próby uczącej w sposób niezależny. Procedura jest następująca¹⁹:

1. Zbiór obserwacji zostaje podzielony na próbę uczącą i próbę testową.
2. Model drzewa powstaje w oparciu o dane z próby uczącej z wykorzystaniem 10-krotnej walidacji krzyżowej, zaś błąd szacowany jest w oparciu o dane z próby testowej, błąd ten oznacza się symbolem e_S .
3. Z próby uczącej losuje się k -podprób bootstrapowych i z każdej z nich buduje jeden model z wykorzystaniem 10-krotnej walidacji krzyżowej, liczba podprób jest zwykle równa 50, co oznacza, że liczba pojedynczych modeli również wynosi 50.
4. 50 pojedynczych drzew zostaje zagregowane w jeden model.
 - 4a. W przypadku drzew klasyfikacyjnych wykorzystuje się tzw. zasadę „głosowania” (*voting*) – rozpoznawany przypadek trafia do tej klasy, do której został przydzielony przez większość z tych 50 drzew, odsetek błędnych klasyfikacji (liczba modeli, w których rozpoznawany obiekt był przydzielany do innej klasy) to błąd predykcji oznaczany symbolem e_B .
 - 4b. W przypadku drzew regresyjnych rozpoznawany przypadek otrzymuje uśrednioną wartość z 50 pojedynczych modeli.
5. Etapy 1-4 powtarza się 100 razy, uśredniając błędy e_S i e_B , w wyniku czego otrzymuje się \bar{e}_S i \bar{e}_B . Wykorzystana tu symbolika bierze się z angielskich terminów *single tree error* (błąd dla pojedynczego drzewa) oraz *bagged tree error* (błąd dla modelu zagregowanego).

Schemat losowania prób i budowy modeli składowych przedstawiono na rys. 4.

¹⁹ L. Breiman, *op. cit.*, 1994, s. 6.



Rys. 4. Schemat losowania prób i budowy modeli składowych w metodzie bagging.

Uważa się, że metoda *bagging* poprawia trafność predykcji w niestabilnych narzędziach analitycznych, za jakie uznano sieci neuronowe, metodę wektorów nośnych oraz drzewa klasyfikacyjne i regresyjne; nie nadaje się natomiast do narzędzi stabilnych, przykładem których jest metoda k-najbliższego sąsiedztwa.

Wykorzystanie wrażliwych algorytmów (*sensitive learning*) – wzmacnianie algorytmów uczących (*boosting*)

Wzmacnianie (*boosting*) to proces służący do poprawy trafności predykcji algorytmów uczących się. Popularnym narzędziem wzmacniającym jest opracowany w 1995 przez Y. Freund'a i R.E. Schapire'a algorytm AdaBoost²⁰. Jego celem, w dużym uproszczeniu, jest stworzenie silnego klasyfikatora poprzez połączenie wielu „słabych” klasyfikatorów. Pod pojęciem klasyfikator²¹ należy rozumieć regułę klasyfikacyjną (zdanie typu „jeżeli ...,

²⁰ Y. Freund, R.E. Schapire, *A Short Introduction to Boosting*, „Journal of Japanese Society for Artificial Intelligence”, September 1999, s. 771-780.

²¹ Autorzy algorytmu używają zamiennie następujących terminów: klasyfikator (*classifier*), hipoteza (*hypothesis*) i uczeń/wynik nauczania (*learner*).



to ...”) wygenerowaną przez dowolne narzędzie analityczne służące do klasyfikacji obiektów, np. drzewo klasyfikacyjne, metodę wektorów nośnych, sieci neuronowe. Terminy „mocny” i „słaby” będą odnosić się odpowiednio do małego błędu klasyfikacji i do dużego błędu klasyfikacji tychże reguł. Pakiet *STATISTICA* zawiera moduł do budowy drzew wzmocnianych (rys. 5).



Rys. 5. Okno opcji drzew wzmocnianych w pakiecie *STATISTICA*.

Akronim AdaBoost można rozwinąć jako adaptacyjne wzmocnianie - od angielskiego terminu *adaptive boosting*. Etapy procedury są tu następujące:

1. Budowa modelu w oparciu o dane z wyjściowego zbioru obserwacji (próby uczącej nr 1), którego rezultatem jest tzw. „słaby” klasyfikator, czyli reguła charakteryzująca się dużym błędem klasyfikacji.
2. Przypisanie większych współczynników wagowych dla przypadków błędnie sklasyfikowanych; powstaje w ten sposób nowy zbiór uczący, na podstawie którego budowany jest kolejny klasyfikator.
3. Etap drugi jest powtarzany za każdym razem, gdy otrzymany model charakteryzuje się dużym odsetkiem błędnych klasyfikacji – stąd też pochodzi nazwa „adaptacyjne wzmocnianie”, algorytm adaptuje się niejako do nowych warunków – do nowego zbioru uczącego z większymi wagami przypadków z klasy mniej licznej.

Podejście mieszane przy użyciu narzędzia Random Forests

Interesującą propozycję rozwiązania problemu nieźrównoważonych klas znaleźć można w pracy Ch. Chena, A. Liawa i L. Breimana z 2004 r.²² Wykorzystano tam metodę losowego lasu²³, łączącą przypisanie większej wagi przypadkom mniej licznym (*sensitive learning*) z redukcją liczebności klasy bardziej licznej (*down-sizing*).

Autorzy pracy zaproponowali tzw. **zrównoważony losowy las** (*balanced random forests*), który tym różni się od pierwowzoru, że używa się w nim warstwowane podpróby bootstrapowe. Procedura postępowania jest następująca:

²² Ch. Chen, A. Liaw, L. Breiman, *Using Random Forest to Learn Imbalanced Data*, Technical Report Nr 666, Department of Statistics, University of California, Berkeley, 2004.

²³ Breiman L., *Random forests*, Statistics Department University of California, Berkeley, 2001.



1. Dla każdego pojedynczego drzewa losuje się podpróbę bootstrapową w ten sposób, że połowa przypadków jest losowana w sposób niezależny z klasy mniej licznej, a druga połowa – również w sposób niezależny – z klasy bardziej licznej.
2. Na podstawie tak dobranych przypadków buduje się model drzewa przy pomocy algorytmu CART bez opcji przycinania; podobnie jak to było w klasycznym losowym lesie, na każdym etapie podziału wykorzystuje się losowo dobrany zestaw predyktorów poddawanych ocenie.
3. Pierwszy i drugi etap powtarza się zadaną liczbę razy, po czym agreguje się pojedyncze drzewa celu uzyskania finalnego modelu predykcyjnego.

Druga modyfikacja losowego lasu to tzw. **ważony losowy las** (*weighted random forests*), w którym wykorzystuje się współczynniki wagowe dla poszczególnych wariantów zmiennej zależnej. Współczynniki te, nazywane również kosztami błędnych klasyfikacji, mają wyższe wartości dla klasy mniej licznej, a niższe dla klasy bardziej licznej. Wykorzystuje się je w dwóch miejscach modelu drzewa klasyfikacyjnego. Najpierw podczas budowy modelu na każdym etapie podziału łączy się je z indeksem Giniego, natomiast później przemnaża się przez nie liczbę przypadków z danej klasy, jaka trafi do węzłów końcowych drzewa klasyfikacyjnego. Koszty błędnych klasyfikacji są istotnym parametrem mającym wpływ na ostateczną postać modelu predykcyjnego. W programie *STATISTICA* można znaleźć moduł umożliwiający budowę losowego lasu (rys. 6).



Rys. 6. Okno opcji losowego lasu w pakiecie *STATISTICA*.

Problem nie zrównoważonych klas jest od dawna utrapieniem analityków zajmujących się budową modeli predykcyjnych. W chwili obecnej wydaje się, że wysiłek powinien zostać skoncentrowany na opracowywaniu wrażliwych algorytmów, wśród których można wymienić chociażby metodę wektorów nośnych. Wzrost mocy obliczeniowej komputerów pozwala łatwiej losować podpróby bootstrapowe, które *de facto* są podstawą wielu istniejących dotychczas narzędzi analitycznych.

Literatura

1. Berry M.J.A., Linoff G.S., *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, Wiley Publishing, 2004.



2. Breiman L. i in., *Classification and Regression Trees*. Chapman and Hall, New York 1984.
3. Breiman L., *Bagging predictors*, Technical Report No. 421, Department of Statistics, University of California, Berkeley, September 1994.
4. Breiman L., *Random forests*, Statistics Department University of California, Berkeley, 2001.
5. Buckinx W., Van den Poel D., *Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting*, Working Paper, Ghent University, May 2003.
6. Chen Ch., Liaw A., Breiman L., *Using Random Forest to Learn Imbalanced Data*, Technical Report Nr 666, Department of Statistics, University of California, Berkeley, 2004.
7. Freund Y., Schapire R.E., *A Short Introduction to Boosting*, „Journal of Japanese Society for Artificial Intelligence”, September 1999, s. 771-780.
8. Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, PWN Warszawa 2001.
9. Łapczyński M., *Data mining w badaniach rynkowych i marketingowych – obszary zastosowań*, Acta Universitatis Lodzianis, Folia Oeconomica 179, 2004, s. 503-509.
10. Lejeune M.A.P.M., *Measuring the impact of data mining on churn management*, Internet Research: Electronic Networking Applications and Policy, Volume 11, Number 5/2001.
11. Mutanen T., *Customer churn analysis – a case study*, Helsinki University of Technology, System analysis Laboratory, Department of Engineering Physics and Mathematics, Independent Research Project in Applied Mathematics, 10 March 2006.
12. Richeldi M., Perrucci A., *Churn Analysis Case Study, Enabling End-User Datawarehouse Mining*, Contract No.: IST-1999-11993, Telecom Italia Lab, December 2002.