Jest to zapis odczytu wygłoszonego na XL Szkole Matematyki Poglądowej, *Matematyczne Obrazki*, styczeń 2008.

Ilustrowana sztuka podejmowania decyzji

Grzegorz HARAŃCZYK, Małgorzata STĘPIEŃ, Kraków

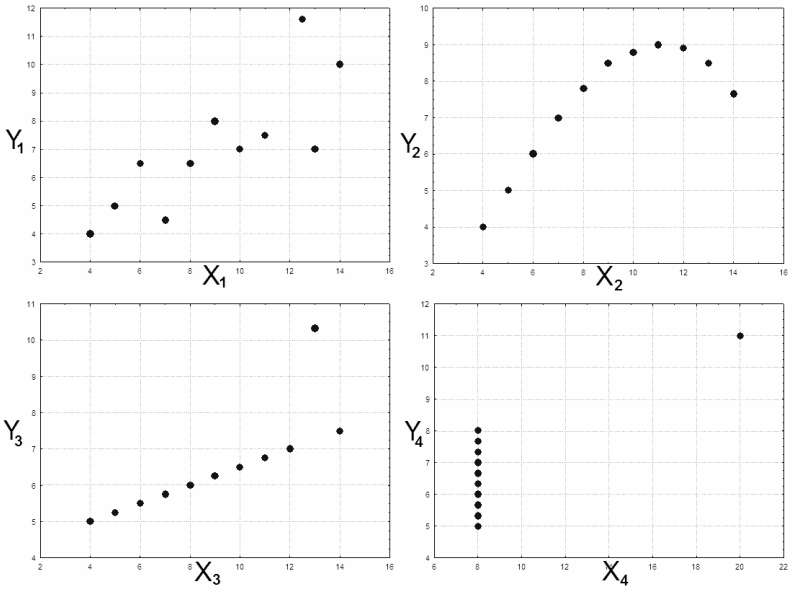
W obecnych czasach dostępnych jest coraz więcej informacji, prawie wszystko jest mierzone, a pomiary archiwizowane, niestety nie przekłada się to w prosty sposób na wiedzę. Znalezienie ukrytych zależności w zbiorze liczb na ogół jest trudnym zadaniem i wymaga odpowiedniego podejścia. Oprócz technik analizy danych i zaawansowanych algorytmów ważną rolę odgrywają również techniki wizualizacji. Wykresy wspomagają statystyczną analizę danych, od analizy danych surowych do oceny zbudowanego modelu [5]. W tym artykule skupimy się na jednej formie wizualizacji przydatnej przy ostatnim etapie analizy.

Wizualizacji danych nie można przecenić, często cytowana jest sentencja amerykańskiego astronoma Edwarda Emersona Barnarda (1857–1923) *jeden wykres jest wart więcej niż setki słów,* jednakowoż często ogranicza się jej rolę do wykonywania tylko prostych zestawień. Dlaczego? Przewrotną tezą, szczególnie w kontekście tytułu tego artykułu, wydawałoby się stwierdzenie, że łatwiej porównywać liczby niż obrazki, ale tak w rzeczywistości jest. Wśród liczb mamy relację porządku z dobrymi własnościami, więc naturalnym jest dążenie do tego, aby na koniec każdego rozumowania wygenerować pewien „magiczny" wskaźnik – jedną liczbę opisującą cały problem, zjawisko. Należy jednak w tej wygodzie zachować umiar, bo często jest zgubna, a wykresy okazują się przydatne.

Angielski matematyk Francis John Anscombe (1918–2001) podał w swojej pracy [1] bardzo przekonywujący przykład na to, że wykresy są niezwykle ważnym elementem analizy danych. Podał cztery zbiory po dwie serie danych: (X_i, Y_i) dla $i = 1, \ldots, 4$ oraz badał współzależność między X_i i Y_i dla $i = 1, 2, 3, 4$. Do badania współzależności między dwiema zmiennymi losowymi na podstawie próby losowej bardzo często wykorzystuje się współczynnik korelacji Pearsona. Wyliczany jest on z bardzo prostej formuły:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Okazuje się, że dla każdego z czterech zbiorów danych otrzymujemy taką samą wartość, mianowicie $r(X_i, Y_i) = 0,8158$ dla $i = 1, 2, 3, 4$. Jest to dość zaskakujące, szczególnie jeśli na wykresie rozrzutu zilustrujemy zależności między badanymi zmiennymi, są zupełnie różne.



Rys. 1. Kwartet Anscombe'a

W dalszej części zostanie omówiona technika krzywych operacyjno-charakterystycznych – krzywych ROC (ang. receiver operating characteristic curve), wykorzystywana w statystycznej teorii decyzji. Podejmowanie decyzji na ogół jest sprawą trudną, jeszcze bardziej komplikuje się właśnie w sytuacji podejmowania decyzji w warunkach niepewności. Typowym zadaniem analizy danych są problemy klasyfikacyjne, polegające na znalezieniu zbioru reguł, tzw. modelu, na podstawie którego można obiekt przyporządkować do jednej z kilku klas. Zakładamy, że mamy g niezależnych prostych prób losowych pochodzących z g różnych populacji (n_i liczność i-tej próby), wszystkie elementy prób losowych są wektorami losowymi o tym samym wymiarze p. Zatem zbiór danych jest postaci:

$$(x_i^{(1)}, \ldots, x_i^{(p)}, y_i), \quad i = 1, \ldots, n_1 + \ldots + n_g,$$

gdzie y_i oznacza etykietę klasy, do której ta obserwacja należy.

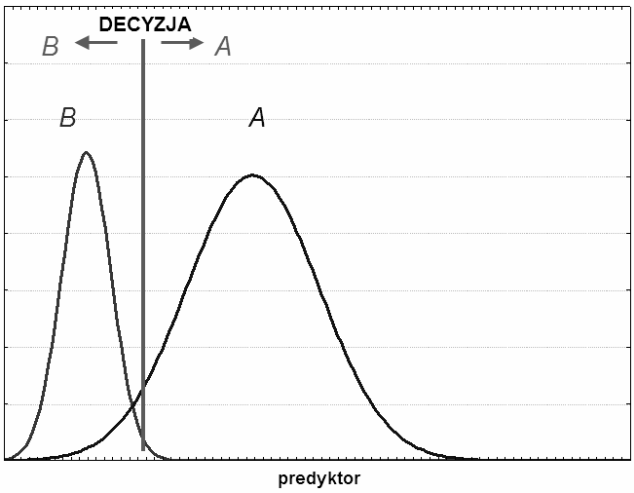
Zadanie klasyfikacji polega na podaniu reguły decyzyjnej przypisującej dowolnej obserwacji $(x_1, \ldots, x_p)$ przynależność do klasy ze zbioru klas $\mathcal{G}$. Innymi słowy mówiąc, należy przewidzieć wartość zmiennej zależnej (określającej klasę), na podstawie wartości pewniej liczby zmiennych niezależnych, nazywanych również predykatorami. Próbę, na podstawie której tworzy się regułę decyzyjną, nazywa się próbą uczącą. Oczywiście w praktyce zależy nam na tym, aby zbudowane reguły działały na całej populacji, z której została wybrana próba ucząca (wdrażanie, stosowanie modelu). Najbardziej znanym tego typu zadaniem jest klasyfikacja kwiatów irysa (por. [2]). Problem ten rozważał Ronald Aylmer Fisher (1890–1962) na zbiorze danych przygotowanym przez amerykańskiego botanika Edgara Andersona (1897–1969). Zbiór ten zawierał charakterystykę 50 kwiatów irysa (kosaćca). Każdy z nich należał do jednej z trzech odmian (Iris setosa, Iris virginica oraz Iris versicolor) i był scharakteryzowany przez długość i szerokość działki oraz długość i szerokość płatka. Zadanie polegało na tym, aby zbudować reguły, które na podstawie tych czterech wielkości charakteryzujących każdy kwiat będą przewidywać, z której odmiany pochodzi.

Obecnie bardzo powszechne jest budowanie modeli klasyfikujących. Stosowane są one między innymi do wykrywania osób należących do grup ryzyka zachorowania na daną chorobę, wykrywania spamu, wytypowania klientów, którzy odpowiedzą na ofertę, czy też tych, którzy chcą zrezygnować z wybranej usługi, rozpoznawanie wzorców i zdefiniowanych obrazów.

W dalszych rozważaniach ograniczymy się do sytuacji, w której zmienna zależna ma tylko dwie wartości – przyporządkowywujemy obiekty do jednej z dwóch klas. Zmiennych niezależnych może być jedna lub więcej (np. w problemie rozważanym przez Fishera były cztery). W sytuacji jednej zmiennej niezależnej problem sprowadza się do przewidywania na podstawie jednej wielkości charakteryzującej dany obiekt, do której z dwóch klas należy. Widać, że ważną kwestią jest tu znalezienie odpowiedniego punktu odcięcia, czyli tej wartości predykatora, powyżej której będziemy decydować, że dany obiekt należy do klasy odpowiadającej większym wartością predykatora. Punkt taki nazywamy punktem odcięcia. Jeśli mamy więcej zmiennych, to na podstawie ich wszystkich chcemy przewidywać przynależność do klasy. Jest wiele metod służących do budowania reguł tego typu, np. drzewa klasyfikacyjne, regresja logistyczna, metoda k-najbliższych sąsiadów, sieci neuronowe. W wyniku zastosowania takiego modelu otrzymujemy prawdopodobieństwo przynależności do wybranej klasy. Przykładowo, w regresji logistycznej budujemy model postaci

$$y = \frac{exp(\sum_i \alpha_i x_i + \beta)}{exp(\sum_i \alpha_i x_i + \beta) + 1}.$$

Jego celem jest przyporządkowanie obiektu do jednej z dwóch klas, jednej kodowanej jako 0, drugiej jako 1, na podstawie wartości $x_1, \ldots, x_n$. Widać, że wartość zmiennej zależnej należy do przedziału $(0, 1)$; oprócz zbudowania modelu (czyli wyznaczenia na postawie danych współczynników $a_1, \ldots, a_n$) ważne jest wybranie odpowiedniego punktu odcięcia, czyli takiej wartości $a \in (0, 1)$, że jeśli $y < a$ to obiekt przyporządkowywujemy do klasy kodowanej przez 0, jeśli $y \geqslant a$



Rys. 2. Punkt odcięcia

to do klasy kodowanej przez 1. Łatwo zauważyć, że przypadek z jednym predykatorem po unormowaniu jego wartości jest szczególnym przypadkiem tej ogólnej sytuacji. Widzimy zatem, że na jednym z etapów problem sprowadza się do szukania optymalnego punktu odcięcia.

Nieodłącznym elementem podejmowania decyzji są błędy (w tym przypadku złe zaklasyfikowania). Błędne decyzje modelu klasyfikującego są nieuniknione, bo często klasy nie są całkowicie separowalne. Łatwo wyobrazić sobie taką sytuację, kiedy dwa obiekty charakteryzowane są za pomocą takich samych wartości zmiennych niezależnych, ale należą do dwóch różnych klas. Z taką sytuacją mamy do czynienia na przykład jeśli nie zebrano odpowiednich danych.

Zestawienie możliwych konsekwencji decyzji widoczne jest poniżej w tabeli. Załóżmy, że dwie klasy zostały zakodowane jako A i B. Wielkości TA, TB, FA, FB to liczby wystąpień odpowiednio poprawnego zaklasyfikowania do klasy A, poprawnego zaklasyfikowania do klasy B, niepoprawnego zaklasyfikowania do grupy A, jeśli obiekt w rzeczywistości należał do grupy B oraz niepoprawnego zaklasyfikowania do grupy B, gdy obiekt pochodził z grupy A.

	Obserwowane A	Obserwowane B
Przewidywane A	TA	FA
Przewidywane B	FB	TB

Dobry model to taki, który minimalizuje liczbę błędów, czyli FB oraz FA. Nie zawsze oba te błędy traktowane są tak samo. W niektórych zastosowaniach te błędne klasyfikacje do dwóch klas mogą mieć bardzo różny koszt. Na przykład, w klasyfikowaniu pacjentów do grup ryzyka gorszym błędem jest traktowanie chorego pacjenta jako zdrowego niż odwrotnie.

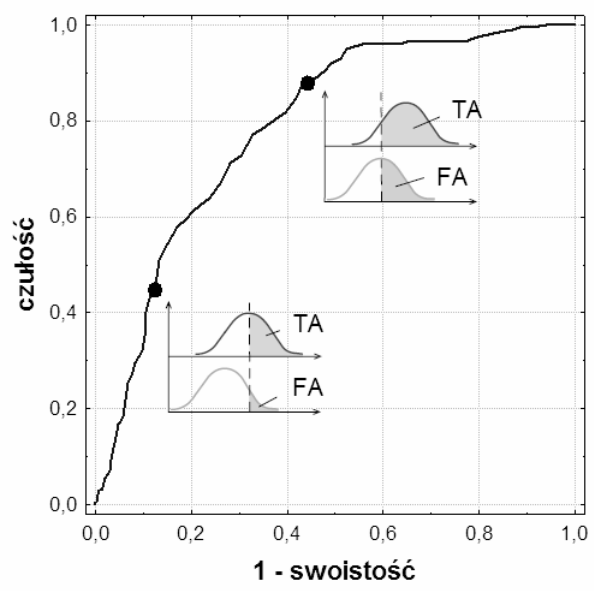
Tak więc znalezienie optymalnego punktu odcięcia ma zapewnić nam najlepsze wyniki – jak najmniejszą liczbę błędów. Wprowadza się dodatkowe kryteria „najlepszych wyników". Jedną z klas przyjmuje się jako wybraną, na ogół tę, której występowanie wiąże się ze wzrostem wartości predykatora. Wielkości, które będziemy chcieli optymalizować podczas procesu decyzyjnego to swoistość (ang. specificity) oraz czułość (ang. sensitivity). Definiujemy je następująco:

$$\text{Czułość} = \frac{TA}{TA + FB},$$

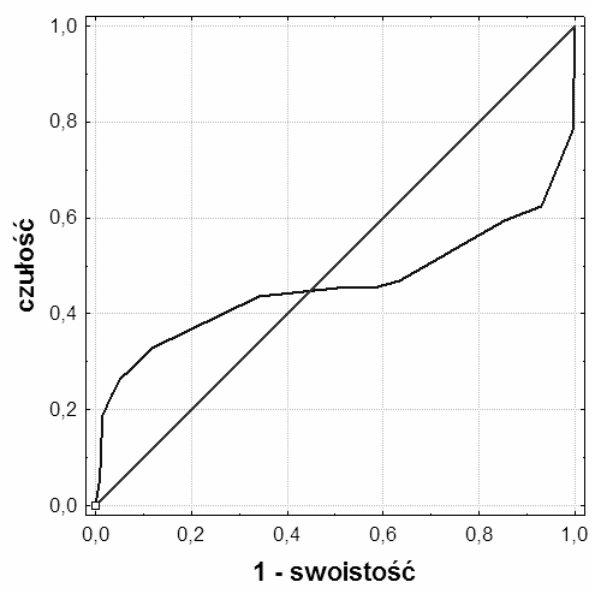
$$\text{Swoistość} = \frac{TB}{TB + FA}.$$

Teoretycznie, czułość określa się jako pole pod krzywą gęstości rozkładu predykatora dla populacji obiektów z wybranej klasy, zaś swoistość określa się analogicznie, jako pole pod krzywą gęstości dla obiektów niewybranej klasy, na prawo od punktu odcięcia. Dobra decyzja to taka, która maksymalizuje obie te wielkości, jednak są to w istocie rzeczy żądania przeciwstawne. Jeśli mamy do czynienia z milionem obserwacji, to mamy milion potencjalnych punktów odcięcia, czyli milion tabel dwa na dwa, a mamy wybrać tę optymalną. Aby dokonać takiego wyboru warto wykorzystać krzywe ROC.

Krzywe ROC zostały wprowadzone podczas II wojny światowej do analizy danych pochodzących z radarów, temu też zawdzięczają swoją nazwę. Ich zadaniem było pomagać operatorom radarów zdecydować, czy zaobserwowany sygnał to wrogi, czy sojuszniczy statek, czy też tylko szum. Po pięćdziesięciu latach krzywe ROC wykorzystywane są w wielu obszarach analizy danych [3, 4], szczególnie popularne są w analizie danych medycznych.



Rys. 3. Konstrukcja krzywej ROC



Rys. 4

Krzywa ROC ilustruje związek miedzy czułością a swoistością dla danego modelu. Dla każdego z możliwych punktów odcięcia obliczamy czułość i swoistość, a następnie zaznaczamy je w układzie współrzędnych, gdzie na osi odciętych jest (1-swoistość), a na osi rzędnych czułość.

Jeśli przyjmujemy równe koszty błędnych klasyfikacji, to optymalnym punktem odcięcia jest punkt krzywej ROC znajdujący się najbliżej punktu o współrzędnych $(0, 1)$. Punkt o współrzędnych $(0, 1)$, to punkt o czułości równej 1 (wszystkie obiekty wybranej klasy wykryto) i swoistości równej 1 (nie uznano błędnie żadnego obiektu za obiekt wybranej klasy). Jeśli dla pewnego punktu odcięcia klasy są całkowicie odseparowane i wskazania modelu dobre, to krzywa ROC przechodzi przez ten punkt. Może się zdarzyć, że model wskazuje klasy „na odwrót" wówczas krzywa ROC przebiega poniżej przekątnej $y = x$. Gdy rozkłady w obu grupach się pokrywają, wówczas krzywa ROC pokrywa się z przekątną $y = x$ (decyzja podejmowana na podstawie modelu jest tak samo dobra jak losowe wybieranie klasy dla danego obiektu).

Oprócz wspomagania wyboru optymalnego punktu odcięcia krzywa ROC używana jest do porównywania różnych modeli, czy to zbudowanych na podstawie różnych zestawów zmiennych niezależnych, czy też różnymi metodami. Zaletą tej metody jest to, że pokazuje siłę wpływu predykatora na występowanie wybranej klasy dla wszystkich możliwych punktów odcięcia. Bardzo popularnym ostatnio podejściem jest wyliczanie pola pod wykresem krzywej ROC, oznaczanego jako AUC. Używane bywa ono jako wskaźnik trafności danego modelu [3]. W ten sposób porównywanie dwóch krzywych często ogranicza się do porównywania jedynie wskaźników AUC, bez wykonywania samych wykresów. Sytuacja staje się podobna do tej z współczynnikiem korelacji opisywanej na początku tego artykułu. Jednak wówczas można przegapić tak ciekawe zjawisko jak to zaprezentowane na poniższym wykresie, nota bene pochodzące z prawdziwych danych. Nietrudno również wygenerować dane dające różne krzywe, dla których pole pod wykresem wynosiłoby 0,8158 . . .

Literatura

[1] F. J. Anscombe *Graphs in Statistical Analysis*, The American Statistician, 1973, 27, 17–21

[2] R. A. Fisher *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics, 1936, 7, 179–188.

[3] J. A. Hanley *Receiver operating characteristic (ROC) methodology: the state of the art* Crit Rev Diagn Imaging. 1989; 29(3), 307–35.

[4] J. A. Swets, R.M. Dawes, J. Monahan *Better decision through science*, Scientific American, 2000, October, 82–87.

[5] E. R. Tufte *The Visual Display of Quantitative Information*, 2nd Edition, Cheshire, CT: Graphics Press, 2001.