



JAKĄ MARCHEW WYBRAĆ NA SURÓWKĘ, CZYLI PRZYKŁAD ZASTOSOWANIA WYBRANYCH MODELI DYSKRYMINACYJNYCH W ANALIZIE DANYCH

Monika Janaszek, Szkoła Główna Gospodarstwa Wiejskiego, Katedra Podstaw Inżynierii

Portret z marchwią w tle

Marchew jest rośliną bardzo cenioną na rynku owocowo-warzywnym, ponieważ stanowi bogate źródło składników odżywczych, ważnych dla ludzkiego organizmu. Ceniona jest również ze względu na obecność w jej korzeniach związków biologicznie czynnych. Wartość odżywczą marchwi określa się na podstawie oceny składu chemicznego, który z kolei jest uwarunkowany genetycznie i różny dla poszczególnych odmian, a wahania zawartości składników mogą być wynikiem zróżnicowania warunków uprawowych (Zadernowski i in., 2003; Nyman i in., 2005). Ponadto jakość pojedynczych korzeni może różnić się od jakości całej partii (Abbot, 1999). Stosowane dotychczas metody oceny wartości odżywczej marchwi zakładają, że korzenie tej samej odmiany, pochodzące z tego samego źródła są wyrównane. Skład chemiczny korzeni jest jednym z najważniejszych determinantów przydatności przetwórczej marchwi. Stosowanie analitycznych metod oceny przetwórczej przydatności marchwi jest jednak kosztowne i czasochłonne, ponadto nie każde przedsiębiorstwo w sektorze przetwórstwa owocowo-warzywnego może pozwolić sobie na utrzymanie zaplecza laboratoryjnego. W praktyce ocena taka jest często przeprowadzana w warunkach niepewnej i niepełnej wiedzy i ogranicza się jedynie do inspekcji pod kątem zdrowotności oraz zawartości substancji szkodliwych i niepożądanych. Informacja o odmianie uznawana jest natomiast za wystarczający determinant przydatności przetwórczej marchwi. Stwarza to trudności i wpływa na precyzję i jednoznaczność tej oceny, a jest ona przecież zasadniczym kryterium, decydującym o możliwości wykorzystania surowca i kierunku jego dalszego przetworzenia (Janaszek, 2008). Potrzebne jest zatem narzędzie umożliwiające zautomatyzowaną ocenę surowca na podstawie informacji, które można zdobyć tanio, łatwo i w krótkim czasie. Takie możliwości dają znane już i szeroko stosowane metody oparte na analizie danych obrazowych. W wyniku analizy obrazu można uzyskać mnóstwo informacji o cechach badanego obiektu, takich jak: kształt, tekstura czy barwa. Systemy wykorzystujące dane obrazowe są już od dawna wdrażane w przemyśle owocowo-warzywnym, np. w sortowniach. O możliwości częściowego lub całkowitego zastąpienia tradycyjnych metod oceny wartości odżywczej marchwi przekonują wyniki badań naukowych. Wykazały one m.in., że różnice



w zawartości cukrów oraz masie suchej substancji i karotenoidów są powiązane z barwą korzeni (Baardseth i in., 1995; Skrede i in., 1997). Horgan i in. (2001) wykazali, że dzięki statystycznej analizie danych obrazowych możliwe jest rozpoznanie odmiany korzeni marchwi na podstawie ich barwy i kształtu.

Wyobraźmy sobie zatem eksperyment, którego celem jest sprawdzenie, czy proste parametry obrazów korzeni, takie jak ich barwa, pozwolą na rozpoznanie grup, różniących się istotnie składem chemicznym (wartością odżywczą). Należy przy tym pamiętać, że marchew, ze względu na anizotropową budowę korzenia i jego morfologiczną niejednorodność, stanowi trudny obiekt badawczy, jeśli chodzi o uzyskanie danych obrazowych, zwłaszcza dotyczących barwy. Prezentowane zagadnienie będzie zatem dotyczyło klasyfikacji wzorcowej korzeni marchwi, na podstawie danych obrazowych, charakteryzujących barwę tych korzeni.

Dyskryminacja niejedno ma imię

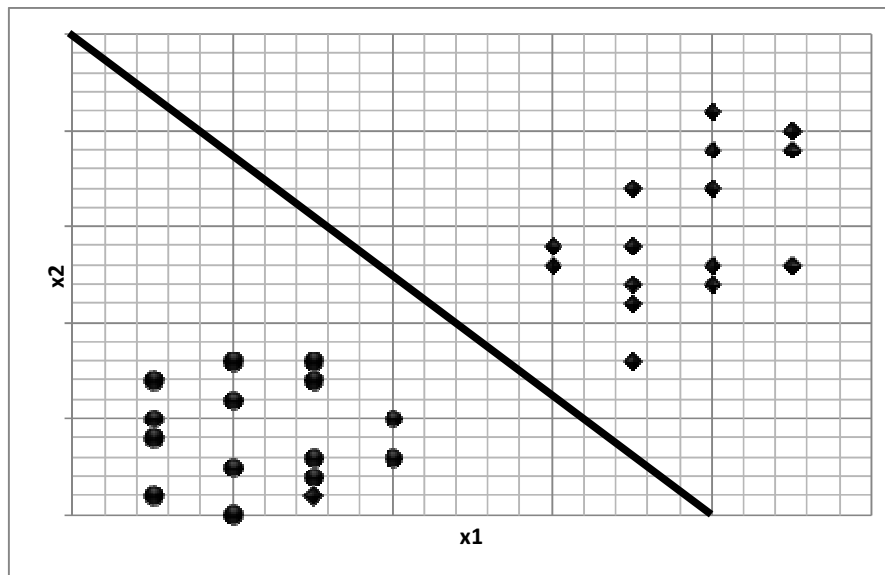
Klasyfikacja wzorcowa czy też uczenie z nauczycielem (nadzorowane) to terminy, którymi określa się analizę dyskryminacyjną. Jej różne nazewnictwo wynika z różnorodności metod, które mogą tę analizę realizować. Można do nich zaliczyć metody klasyczne, metodę k -najbliższych sąsiadów, sieci neuronowe oraz metodę wektorów wspierających.

Z matematycznego punktu widzenia zagadnienie rozważane w analizie dyskryminacyjnej polega na zaklasyfikowaniu obiektów (obserwacji) o nieznanym pochodzeniu do jednej z dwóch lub więcej znanych populacji (klas, skupień), na podstawie wartości obserwowanych cech tych obiektów (Krusińska, 1987). Kryterium klasyfikacji obiektów do określonych populacji stanowi wartość funkcji dyskryminacyjnej, a postać tej funkcji ustala się na podstawie uprzednio sklasyfikowanych danych historycznych. Klasyczne podejście do analizy funkcji dyskryminacji opiera się na bayesowskiej teorii prawdopodobieństwa. W podejściu tym zakłada się, że znane są prawdopodobieństwa przynależności do rozpatrywanych klas. Prawdopodobieństwo to nazywa się prawdopodobieństwem *a priori* i najczęściej jako jego estymator wykorzystuje się frakcje obiektów należących do poszczególnych klas. Prawdopodobieństwo *a priori* można również określić subiektywnie lub przyjąć, że jest ono jednakowe we wszystkich klasach (Gatnar, 2009). Zgodnie z twierdzeniem Bayesa o prawdopodobieństwie warunkowym można oszacować prawdopodobieństwo *a posteriori* przynależności obiektu do konkretnej klasy. Odpowiada to wydzieleniu w n -wymiarowej przestrzeni wektorów zmiennych opisujących obiekty takiej podprzestrzeni, w której prawdopodobieństwo *a posteriori* dla obiektu jest największe.

Nie należy mylić analizy dyskryminacyjnej z analizą skupień, mimo iż pojęcia „klasa” i „skupienie” stosuje się w obu tych analizach zamiennie. Wszystkie odmiany analizy dyskryminacyjnej wymagają uprzedniej wiedzy na temat klas, zazwyczaj w postaci próbek z każdej z nich. W analizie skupień dane nie obejmują informacji na temat przynależności klasowej, a celem tej analizy jest samodzielne przeprowadzenie klasyfikacji poprzez rozpoznanie struktury danych.

Termin „analiza dyskryminacyjna” odnosi się w rzeczywistości do kilku różnych typów analizy. Jedną z nich jest klasyfikacyjna analiza dyskryminacyjna (ang. *Classificatory Discriminant Analysis*). Głównym jej celem jest znalezienie matematycznej reguły lub funkcji dyskryminacyjnej, która określa przynależność obserwacji do określonej, znanej klasy, na podstawie opisujących ją zmiennych ilościowych. Najprostszym przypadkiem tej analizy jest tzw. test dychotomiczny, polegający na wyznaczeniu wartości krytycznej, rozdzielającej obiekty na dwie klasy tylko na podstawie wartości jednej zmiennej. W klasyfikacyjnej analizie dyskryminacyjnej istnieją dwa podejścia: parametryczne, w którym zakładamy, że zmienne mają wewnątrz klas rozkład zbliżony do wielocechowego rozkładu normalnego, oraz nieparametryczne, jeżeli powyższe założenie nie jest spełnione lub nic nie wiemy na temat rozkładu zmiennych.

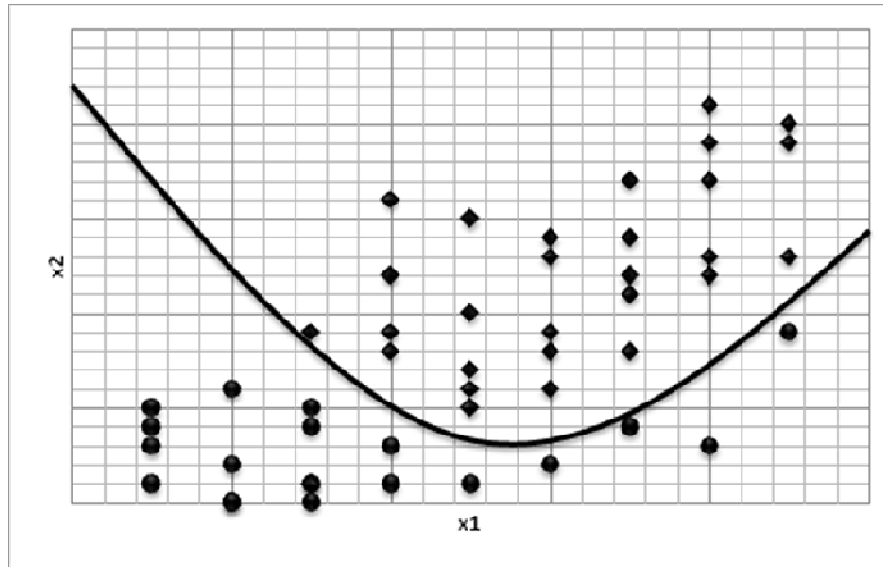
Metody parametryczne umożliwiają wyznaczenie liniowej reguły dyskryminacji (ang. *LDF – Linear Discriminant Function*), o ile spełnione jest założenie o tym, że macierze kowariancji wyznaczone dla rozpatrywanych klas są równe. Warunek ten można sprawdzić za pomocą zmodyfikowanego testu Bartletta na homogeniczność kowariancji wewnątrzgrupowych. Jeśli warunek jest spełniony, obiekty powinny tworzyć skupienia w postaci gniazd, a w ogólnym przypadku w postaci hipersfer tej samej wielkości (Morrison, 1976; Anderson, 1984). Liniową funkcję dyskryminacyjną można wyznaczyć dla każdej klasy, a obiekt klasyfikowany trafi do skupienia, w którym wartość tej funkcji będzie największa. W najprostszym przypadku wyznaczenie reguły klasyfikacji odpowiada określeniu położenia prostej rozdzielającej klasy na płaszczyźnie (rys. 1), ogólnie zaś położeniu hiperpłaszczyzny rozdzielającej klasy w przestrzeni n -wymiarowej.



Rys. 1. Liniowa funkcja dyskryminacyjna.

Jeśli macierz kowariancji jest diagonalna, wówczas hiperpłaszczyzna rozdzielająca dwa skupienia powinna być prostopadła do linii łączącej ich środki ciężkości. Obiekt będzie przydzielony do klasy, której środek ciężkości leży najbliżej w sensie odległości euklidesowej. Jeżeli macierz kowariancji nie jest diagonalna, to o przynależności obiektu do

skupienia decydować będzie odległość D^2 Mahalanobisa od środka ciężkości skupienia. Jeżeli macierze kowariancji dla klas są różne, zamiast funkcji liniowej wyznacza się kwadratową funkcję dyskryminacji (ang. *QDF – Quadratic Discriminant Function*). W wielu przypadkach funkcja kwadratowa (rys. 2) jest jednak mniej praktyczna i trudno interpretowalna ze względu na swoją zawiłą postać. Daje wyraźnie gorsze wyniki w porównaniu do funkcji liniowej i może okazać się niestabilna przy klasyfikacji nowych przypadków.



Rys. 2. Kwadratowa funkcja dyskryminacyjna.

W podejściu nieparametrycznym kryterium klasyfikacji można poszukiwać metodą k -najbliższych sąsiadów (ang. *KNN – k-Nearest Neighbours*). Metoda KNN polega na klasyfikowaniu obiektu do zbioru na podstawie otoczenia tego obiektu. Rozpatrywany przypadek będzie należał do tej grupy, do której należy większość jego sąsiadów. Obszar sąsiedztwa zwykle definiuje się przez wyznaczenie promienia hipersfery, ze środkiem w rozpatrywanym punkcie x_i . Metoda ta ma swoje wady: są to m.in. uznaniowość przy ustalaniu promienia hipersfery czy miary odległości między obiektami oraz niska wydajność w przypadku dużych zbiorów danych, co z kolei wiąże się z koniecznością stosowania dodatkowych algorytmów ograniczających liczbę obiektów (algorytmy zagęszczania, redukcji lub edycji zbioru danych). Do szacowania gęstości prawdopodobieństwa przynależności obiektu do klasy można również wykorzystać estymatory jądrowe (ang. *kernel functions*). Estymatorem jądrowym $\hat{f} : \mathbb{R}^n \rightarrow [0, \infty)$ funkcji gęstości f , n -wymiarowej zmiennej losowej X , nazywamy wyrażenie:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^m K\left(\frac{x - x_i}{h}\right),$$

w którym $m \in \mathbb{N} \setminus \{0\}$ oznacza licznosc próbki, $h > 0$ jest parametrem wygładzania, a $K : \mathbb{R}^n \rightarrow [0, \infty)$ nazywane jądrem jest funkcją symetryczną względem zera i mającą w tym punkcie słabe maksimum. W analizie dyskryminacyjnej do najpopularniejszych



jąder należą: jądro normalne, Epanechnikova, dwuwagowe, trójkątne i jednostajne (Kulczycki, 2005).

Gdzie kucharek sześć...

W przypadku gdy analizie podlegają obiekty opisane dużą liczbą zmiennych, klasyfikacyjna analiza dyskryminacyjna może nie poradzić sobie z nadmiarem informacji i dyskryminacja nie da dobrego efektu. W takiej sytuacji warto ograniczyć liczbę zmiennych w pierwotnej przestrzeni analizy do tych, które mają największą zdolność dyskryminacyjną. Takie działanie nazywamy krokową analizą dyskryminacyjną (ang. *Stepwise Discriminant Analysis*) i może ona zostać przeprowadzona „w przód”, „w tył” i w obu kierunkach. Selekcja „w przód” rozpoczyna się od modelu, który nie zawiera żadnych zmiennych (nie ma zdolności dyskryminacyjnej). W kolejnych krokach do modelu wprowadza się po jednej zmiennej, aż do momentu, gdy żadna ze zmiennych poza modelem nie spełnia kryterium wejścia. W przypadku selekcji „w tył” w modelu znajdują się wszystkie zmienne. W kolejnych krokach usuwa się po jednej zmiennej, aż do chwili, gdy żadna ze zmiennych pozostających w modelu nie spełnia kryterium wyjścia. Selekcja dwukierunkowa również rozpoczyna się od modelu niemającego zdolności dyskryminacyjnej. W każdym kroku do modelu wprowadza się zmienną, która w największym stopniu spełnia kryterium wejścia i usuwa się zmienną, która w największym stopniu spełnia kryterium wyjścia. Jeżeli żadna ze zmiennych poza modelem nie może zostać do niego wprowadzona, ani też żadna ze zmiennych w modelu nie może zostać z niego usunięta, dobór zmiennych jest zakończony. W przypadku każdej procedury krokowego doboru zmiennych wykonywanych jest wiele testów istotności, każdy na założonym poziomie (zwykle jest to 0,05). Należy pamiętać, że ogólne prawdopodobieństwo popełnienia błędu pierwszego rodzaju, czyli odrzucenia co najmniej jednej hipotezy prawdziwej, jest w tym przypadku znacznie większe niż założony przez nas poziom istotności. Logiczne więc wydaje się, że ustalanie istotności na bardzo niskim poziomie powinno zapobiec włączaniu i pozostawianiu w modelu zmiennych o niskiej zdolności dyskryminacyjnej. Tymczasem Costanza i Afifi (1979), badając różne kryteria zatrzymania selekcji „w przód”, wykazali, że umiarkowane wartości poziomów istotności (rzędu 0,1 do 0,25) częściej dają lepsze rezultaty niż poziomy bardzo niskie lub bardzo wysokie.

Z nadmiarem zmiennych w analizie można poradzić sobie również w inny sposób. Wszystkie opisane dotychczas działania miały na celu odnalezienie reguły klasyfikacji, aby dobrać funkcje dyskryminacyjne w pierwotnej przestrzeni analizy. Inne podejście polega na znalezieniu takich liniowych kombinacji zmiennych, które zapewnią kolejnym funkcjom dyskryminacyjnym optymalną i coraz bardziej drobiazgową klasyfikację obiektów. A zatem należy zastąpić oryginalne cechy zmiennymi kanonicznymi (tzw. nieelementarnymi zmiennymi dyskryminacyjnymi). Zmienne kanoniczne to liniowe kombinacje cech oryginalnych, które niosą maksimum obserwowanej zmienności danych źródłowych (maksimum informacji zapisanych w cechach źródłowych). Kanoniczna analiza dyskryminacyjna (ang. *Canonical Discriminant Analysis*) jest techniką redukcji wymiarów, a jej zapleczem

teoretycznym jest metoda składowych głównych i analiza korelacji kanonicznych, zaproponowana przez Fishera (1936).

Kuchnia współczesna

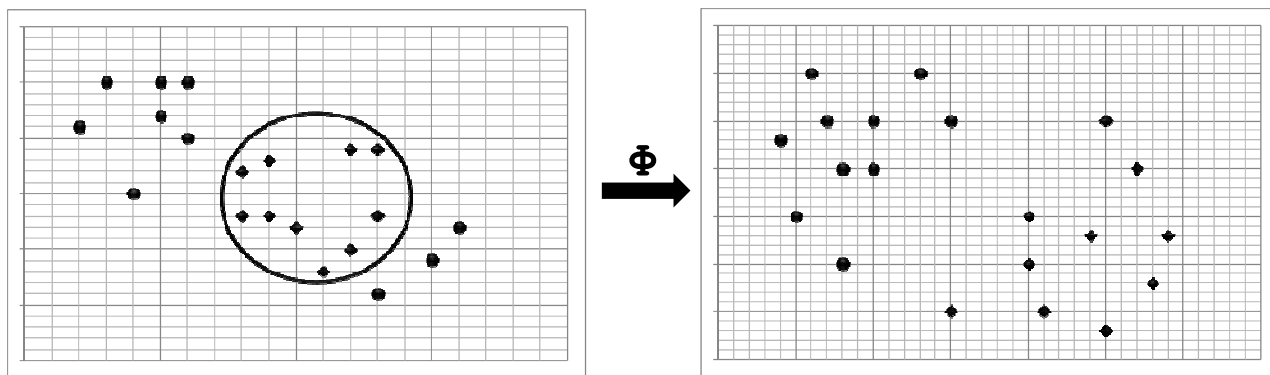
Vapnik (1995) zaproponował odmienne podejście do analizy dyskryminacyjnej i nazwał je metodą wektorów nośnych (ang. *SVM – Support Vector Machine*). Metoda ta może być realizowana przez sieć neuronową, wykorzystującą znane algorytmy uczenia nadzorowanego, przydatne w zagadnieniach rozpoznawania wzorców i regresji. Koncepcja metody SVM opiera się na podziale przestrzeni decyzyjnej poprzez określanie granic liniowo separujących obiekty o różnej przynależności klasowej. W metodzie tej poszukujemy hiperpłaszczyzny określonej ogólnie wzorem:

$$Y = \alpha_0 + \boldsymbol{\alpha} \mathbf{X}^T,$$

gdzie $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_m]$ jest wektorem parametrów, prostopadłym do tej hiperpłaszczyzny. Jeśli obiekty nie są liniowo separowalne, wówczas następuje transformacja zbioru danych do przestrzeni o dużo większym wymiarze, w której separacja liniowa klas jest już możliwa (rys. 3). Transformację tę ogólnie zapisujemy:

$$\Phi = X^m \rightarrow Z,$$

co oznacza, że obserwacji $[x_i, y_i]$ w przestrzeni X^m odpowiada obserwacja $[\varphi(x_i), y_i]$ w przestrzeni Z .



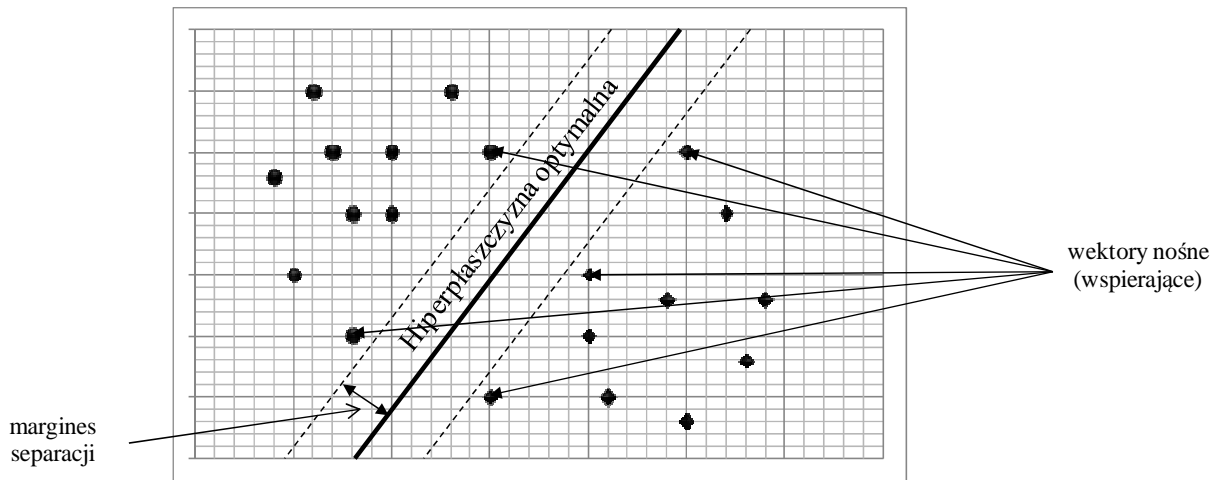
Rys. 3. Idea SVM.

Transformacja ta jest najczęściej nieliniowa, ale postać funkcji transformującej nie musi być znana, wystarczy znajomość iloczynu skalarnego:

$$K(\mathbf{u}, \mathbf{v}) = \boldsymbol{\varphi}(\mathbf{u}) \cdot \boldsymbol{\varphi}^T(\mathbf{v}),$$

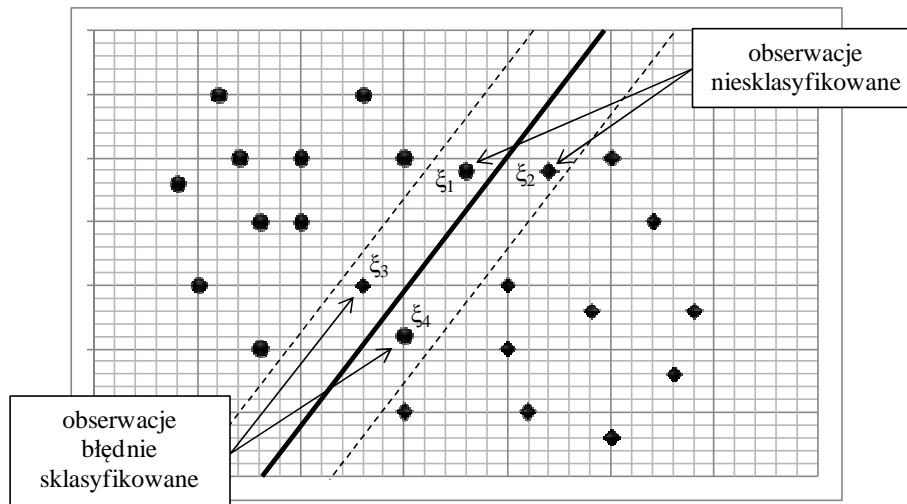
gdzie \mathbf{u} , \mathbf{v} to wektory, a za $K(\mathbf{u}, \mathbf{v})$ przyjmuje się zwykle funkcję jądra. Do najpopularniejszych funkcji jądra, wykorzystywanych w metodzie SVM, należą: funkcja Gaussa, wielomianowa, sigmoidalna i liniowa. Z praktyki wynika, że najlepsze rezultaty uzyskuje się dla funkcji gaussowskiej.

Istotą metody SVM jest taki dobór parametrów, aby obszar, w którym znajdzie się hiperpłaszczyzna, był możliwie najszerszy. Otoczenie obserwacji leżących najbliżej hiperpłaszczyzny nazywa się marginesem (rys. 4). Szeroki margines będzie zapewniał prawidłowe (optimalne) położenie hiperpłaszczyzny separującej klasy. Rozwiązanie polega zatem na odnalezieniu takiego wektora parametrów $\alpha = [\alpha_1, \dots, \alpha_m]$, który będzie maksymalizował wielkość marginesu, czyli takiego, że minimalny iloczyn zmiennej zależnej i odległości skierowanej od hiperpłaszczyzny do wszystkich obserwacji w zbiorze będzie największy (Vapnik, 1998). Oznacza to minimalizację normy wektora alfa. Do rozwiązania tego zadania optymalizacyjnego wykorzystuje się mnożniki Lagrange'a. W rzeczywistości do optymalizacji położenia hiperpłaszczyzny nie potrzeba wszystkich obserwacji.



Rys. 4. Margines separacji klas.

Część z nich zwykle leży poza obszarem, w którym mogłaby się znaleźć hiperpłaszczyzna. Obserwacje wykorzystywane do optymalizacji, a więc takie, które określają szerokość marginesu, ponieważ leżą najbliżej niego, noszą nazwę wektorów nośnych lub wspierających.



Rys. 5. Parametryzacja funkcji kryterium.

W przypadku braku liniowej separowalności klas w obszarze marginesu mogą znaleźć się obserwacje, które nie zostały przydzielone do żadnej klasy. Zdarza się również, że obserwacje znajdują się po przeciwnej stronie hiperpłaszczyzny, co oznacza że są błędnie sklasyfikowane. Do modelu dyskryminacyjnego wprowadza się wówczas dodatkowe zmienne (przesunięcia). Wnoszą one informację o tym, jak silna jest tendencja obserwacji do pozostawania w obszarze marginesu lub w jakim stopniu może ona zostać błędnie sklasyfikowana (rys. 5).

Aby zapewnić funkcji dyskryminacyjnej możliwie największą dokładność, należy ograniczyć liczbę takich obserwacji. Na przesunięcia nakłada się zatem ograniczenia, że powinny być one dodatnie lub sumować się do określonej wartości, a do funkcji kryterium wprowadza się parametr regularyzacji C :

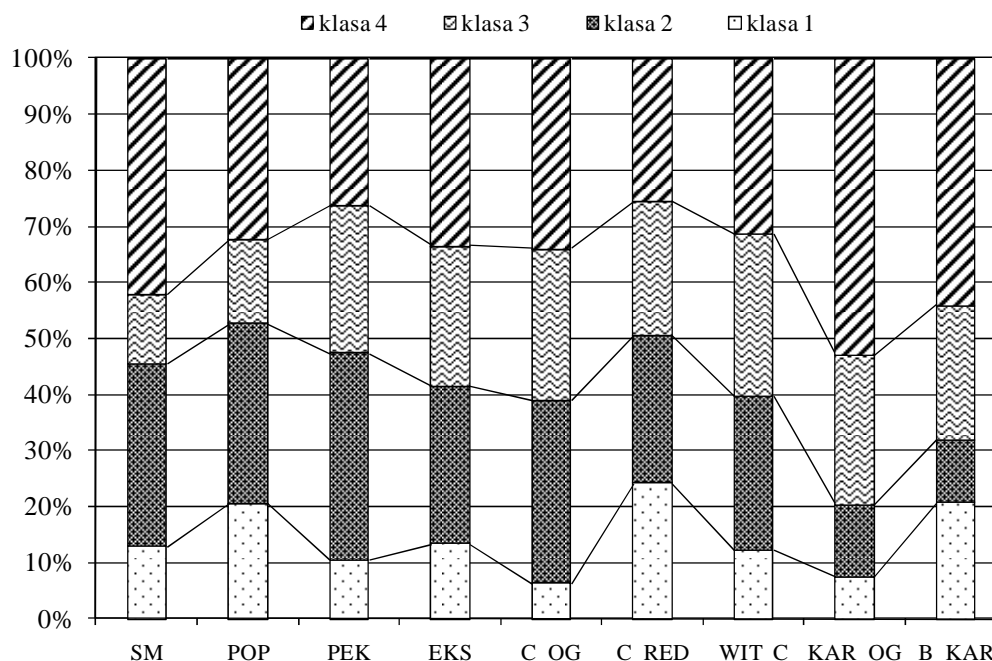
$$Y(\alpha \mathbf{X}^T + \alpha_0) \geq C(1 - \xi_i) \wedge \sum_i \xi_i = const$$

Jego zadaniem jest zapewnienie równowagi między zdolnością sieci do aproksymacji a zdolnościami do uogólniania funkcji dyskryminacyjnej. W takim przypadku na położenie hiperpłaszczyzny wpływa wyłącznie parametr C , którego wartość zwykle ustala się empirycznie lub w tzw. sprawdzanie krzyżowym. Zależność między parametrem C a szerokością marginesu jest odwrotna (Gatnar, 2009).

Składniki i receptury

W doświadczeniu wykorzystano korzenie marchwi jadalnej (*Daucus carota* L.) odmian: Florida, Kathmandu, Kazan, Laguna, Mazurska, Recoleta, Sugarsnax, pochodzące z dwóch kolejnych lat uprawy na plantacji RZD w Żelaznej. W każdym roku badań wykonano analizę składu chemicznego korzeni, oznaczając: masę suchej substancji (SM), popiół (POP), zawartość pektyn (PEK), ekstrakt ogólny (EKS), cukry ogółem (C_OG) i redukujące (C_RED), witaminę C (WIT_C), karotenoidy ogółem (KAR_OG) i β -karoten (B_KAR). Parametry barwy korzeni marchwi otrzymano z obrazów ich przekrojów wzdłużnych, utrwalonych aparatem cyfrowym OLYMPUS 5050Z. Wszystkie zdjęcia wykonano w kolorze, przy czułości 100 ISO, ogniskowej 21,3 mm (odpowiednik ogniskowej 105 mm w aparacie małoobrazkowym 35 mm) i przysłonie f2,8. Korzenie umieszczano na powierzchni, wykonanej z białego, przepuszczającego i równomiernie rozpraszającego światło poliwęglanu. Powierzchnia fotografowanego obiektu oświetlana była źródłem światła o temperaturze barwowej 5000 K. Przed wykonaniem zdjęć wykonywano kalibrację aparatu, poprzez ustawienie balansu bieli. Na podstawie wartości składowych R, G, B i L, a, b, określonych dla każdego piksela obrazu, wyznaczono wektory średnich tych składowych dla każdego korzenia, a następnie przeskalowano do wartości z zakresu [0, 1]. Określono długość i szerokość każdego korzenia oraz ich walca osiowego, zgodnie z metodyką opisaną przez Horgana (2001). Wykonane pomiary wykorzystano do ekstrapolacji objętości korzeni i ich walców osiowych przy założeniu stożkowatego kształtu korzeni. Udział rdzenia (UR) wyrażono jako stosunek objętości walca osiowego do objętości całego korzenia. Ze względu na brak jasno sprecyzowanych i sformalizowanych wymagań co do

cech surowca przeznaczanego na określone cele przetwórcze, klasyfikację poprzedzono wyodrębnieniem grup obiektów o podobnym składzie chemicznym (rys. 6). Pojedyncza grupa (skupienie) stanowiła klasę obiektów, którą wykorzystano jako wzorzec w analizach dyskryminacyjnych. Zbiór danych zawierał zmienną grupującą (wzorzec) oraz siedem zmiennych (deskryptorów) opisujących klasyfikowane obiekty. W nawiązaniu do wyników badań nad przydatnością przetwórczą warzyw i owoców (Ibarz i in., 1999; Krokida i in. 2001), a w szczególności marchwi (Talcott i in., 2001; Borowska i in., 2005; Zielińska i in., 2005) udało się dokonać wstępnej charakterystyki dwóch z czterech wyłonionych klas. Klasa pierwsza charakteryzowała się niskimi zawartościami oznaczonych substancji, co raczej nie czyni jej przydatną do przetwórstwa, głównie ze względu na niską masę suchej substancji oraz najmniejszą zawartość cukrów ogółem i karotenoidów. Ze względu na wysoką (lecz porównywalną do pozostałych klas) zawartość cukrów redukujących nie jest to dobry surowiec na susz. Klasa druga odpowiadała marchwi przydatnej do produkcji soków przecierowych. Pożądane cechy w tym przypadku to: wysoki ekstrakt, masa suchej substancji i zawartość cukrów ogółem. Klasa trzecia, mimo niskiej masy suchej substancji, mogłaby być scharakteryzowana jako surowiec do produkcji mrozonek. Do cech pożądanych w tym przypadku należą: wysoki ekstrakt, zawartość cukrów ogółem, pektyn i karotenoidów, a korzenie w tej klasie charakteryzują się przeciętną zawartością tych substancji.



Rys. 6. Charakterystyka klas (wzorców SVM) pod względem składu chemicznego.

Klasa czwarta odpowiadała raczej marchwi pożądanej w produkcji soków zagęszczonych, gdzie surowiec powinien charakteryzować się wysokim ekstraktem, zawartością cukrów i karotenoidów.

W przypadku rozpatrywanego zagadnienia analizę dyskryminacyjną przeprowadzono metodą ni-SVM (ν -SVM), z radialnym typem jądra (RBF). Dane wprowadzane do sieci podzielono na dwa zbiory: uczący i testowy. Zbiory te zawierały odpowiednio 75% i 25%



obserwacji. Na wejścia sieci wprowadzono ciąg uczący, w którego skład weszły wartości zmiennych R, G, B, L, a, b i UR. Uczenie modelu SVM poprzedzono dziesięciokrotną walidacją krzyżową, w celu doboru optymalnych wartości parametrów ν oraz γ . Proces uczenia sieci obejmował tysiąc iteracji, a warunkiem zatrzymania uczenia było osiągnięcie błędu na poziomie 0,001.

Surówka gotowa

Zadaniem sieci SVM była prawidłowa klasyfikacja korzeni marchwi do wyodrębnionych wcześniej skupień, ale tylko na podstawie danych o ich barwie. Po dziesięciokrotnej walidacji krzyżowej dobrano dla sieci SVM parametry ν oraz γ , których wartości wyniosły odpowiednio 0,4 i 0,1. Jakość działania sieci oceniono na podstawie błędów, z jakimi rozpoznała ona przynależność obiektów ze zbioru testowego do zdefiniowanych wcześniej klas. Trafność klasyfikacji w zbiorze uczącym wyniosła 91,43%, a w zbiorze testowym 94,29%. Liczba wektorów związanych wyniosła 7. Procentowe udziały przypadków poprawnie przypisanych do wzorców były następujące: skupienie 1 – 96,67%, skupienie 2 – 88,89%, skupienie 3 – 90%, skupienie 4 – 100%. Wysoka trafność klasyfikacji w przypadku skupienia 4 wynikała z małej liczności tego skupienia. Tego typu sytuacji można w przyszłości zapobiegać przez stosowanie tzw. przepróbkowania w procesie wyłaniania zbioru testowego.

Podsumowanie

Wyniki klasyfikacji wskazują, że informacja o barwie wprowadzona do sieci w postaci wektora średnich nie była informacją wystarczającą do tego, aby bardzo precyzyjnie klasyfikować wyłonione wcześniej wzorce. Mimo to uzyskana trafność klasyfikacji przekonuje, iż na podstawie barwy można odwzorować korzenie marchwi charakteryzujące się podobnymi cechami składu chemicznego. Weryfikuje to wysuniętą hipotezę, że istnieją zależności między barwą a składem chemicznym korzeni marchwi, oraz przekonuje, że analiza składu chemicznego oraz barwy może być wykorzystana do opracowania wielokryterialnej klasyfikacji marchwi pod względem jej przydatności przetwórczej. Należy również przypuszczać, że tego typu klasyfikacja nie będzie oparta na zależnościach między pojedynczymi cechami składu chemicznego i składowymi barwy. Barwa stanowi bowiem odwzorowanie ogólnej informacji o składzie chemicznym korzeni, a nie o zawartości konkretnej substancji (np. karotenoidów). Zastosowanie sieci SVM w zagadnieniu klasyfikacji korzeni marchwi potwierdziło jej wysoką skuteczność, o czym świadczy trafność klasyfikacji w zbiorze testowym. Przydatność sieci SVM w rozwiązywaniu problemów klasyfikacji wynika z samodzielnego i optymalnego doboru struktury sieci, zapewniającego zdolności uogólniające. Ograniczona liczba wektorów podtrzymujących, a tym samym zredukowana struktura sieci, umożliwi skuteczną jej zastosowanie w mniej licznych zbiorach.



Literatura

1. Abbott J. A. 1999. Quality measurement of fruits and vegetables. *Postharvest Biology and Technology*, 15: 207–225.
2. Anderson T. W. 1984. *An introduction to multivariate statistical analysis*, 2nd edition. John Wiley & Sons Inc., New York.
3. Baardseth P., Rosenfeld H. J., Sundt T. W., Skrede G., Lea P., Slinde E. 1995. Evaluation of carrot varieties for production of deep-fried carrot chips–I. Chemical aspects. *Food Research International*, 28 (3): 195–200.
4. Borowska E. J., Zadernowski R., Szajdek A., Majewska K., Budrewicz G. 2005. Cechy organoleptyczne, fizyczne i chemiczne wybranych odmian marchwi przydatnych do produkcji soku. *Polish Journal of Natural Sciences*, 18 (1): 174–186.
5. Constanza M. C., Afifi A. A. 1979. Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association*, 74: 777–785.
6. Fisher R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7: 179–188.
7. Gatnar E. 2009. Analiza dyskryminacyjna. W: Walesiak M., Gatnar E. (red.). *Statystyczna analiza danych z wykorzystaniem programu R*. PWN, Warszawa, 193–237.
8. Horgan G. W. 2001. The statistical analysis of plant part appearance – a review. *Computers and Electronics in Agriculture*, 31: 169–90.
9. Horgan G. W., Talbot M., Davey J. C. 2001. Use of statistical image analysis to discriminate carrot cultivars. *Computers and Electronics in Agriculture*, 31: 191–199.
10. Ibarz A., Pagán J., Garza S. 1999. Kinetic models for colour changes in pear puree during heating at relatively high temperatures. *Journal of Food Engineering*, 39: 415–422.
11. Janaszek M. A. 2008. Identyfikacja cech korzeni marchwi jadalnej z wykorzystaniem komputerowej analizy obrazu. *Rozprawa doktorska*. SGGW, Warszawa.
12. Krokida M. K., Maroulis Z. B., Saravacos G. D. 2001. The effect of the method of drying on the colour of dehydrated products. *International Journal of Food Science and Technology*, 36: 53–59.
13. Krusińska E. 1987. Wybrane metody analizy dyskryminacyjnej i ich zastosowanie do wspomagania diagnozy lekarskiej. *Człowiek, Populacja, Środowisko - Prace Dolnośląskiego Centrum Diagnostyki Medycznej DOLMED we Wrocławiu*, 4 (23): 243–283.
14. Kulczycki P. 2005. *Estymatory jądrowe w analizie systemowej*. WNT, Warszawa.
15. Morrison D. F. 1976. *Multivariate statistical methods*. McGraw-Hill, New York.
16. Nyman M. E. G. L., Svanberg S. J. M., Andersson R. & Nilsson T. (2005). Effects of cultivar, root weight, storage and boiling on carbohydrate content in carrots (*Daucus carota* L.). *Journal of the Science of Food and Agriculture*, 85, 441–449.
17. Skrede G., Nilsson A., Baardseth P., Rosenfeld H. J., Enersen G., Slinde E. 1997. Evaluation of carrot varieties for production of deep-fried carrot chips–III. Carotenoids. *Food Research International*, 30 (1):73–81.



18. Talcott S. T., Howard L. R., Brenes C. H. 2001. Factors contributing to taste and quality of commercially processed strained carrots. *Food Research International*, 34: 31–38.
19. Vapnik V. N. 1995. *Nature of Statistical Learning Theory*. Springer.
20. Vapnik V. N. 1998. *Statistical Learning Theory*. Wiley, New York.
21. Zadernowski R., Budrewicz G., Borowska E. J., Kaszubski W. 2003. Sok z marchwi naturalnie mętny – kryteria doboru surowca oraz optymalizacji procesu technologicznego (1). *Przemysł Fermentacyjny i Owocowo-Warzywny*, 5: 15–16.
22. Zielińska M., Zapotoczny P., Markowski M. 2005. Colour standard and homogenous groups of dried carrots of 34 commercial varieties. *Polish Journal of Food and Nutrition Sciences*, 14/55 (1): 51–56.