**Selekcja informacji w eksploracji danych**

**z wykorzystaniem programu RapidMiner**

**Streszczenie**

Przedstawiony zostanie przeglądowo aktualny stan wiedzy w zakresie selekcji informacji
w eksploracji danych, ze szczególnym uwzględnieniem najnowszych metod, a w szczególności zagadnienie selekcji cech i selekcji wektorów oraz integracja selekcji cech z selekcją wektorów
w zastosowaniu do redukcji rozmiaru danych i eliminacji szumu w problemach klasyfikacji
i regresji. Zostaną przeprowadzone praktyczne demonstracje w programie RapidMiner oraz pokazane jak moduły RapidMinera można uruchomić z poziomu własnej aplikacji. Tak stworzona prezentacja umożliwi zarówno zapoznanie sięz zagadnieniami selekcji informacji, jaki
z posługiwaniem się programem RapidMiner wrazz możliwością wykorzystania go we własnym systemie informatycznym.

**Wprowadzenie**

Zagadnienia eksploracji danych występują często w wielu gałęziach przemysłu, począwszy od inżynierii materiałowej, gdzie są różne dane pomiarowe procesów technologicznych, poprzez medycynę – dane z badań pacjentów, poprzez dane o zachowaniach klientów (banków, linii lotniczych) aż do danych z rynków finansowych. Celem eksploracji takich danych jest podjęcie odpowiednich decyzji prowadzących do poprawy jakości produkcji/leczenia/obsługi klientów
i zwiększenie zysków. Przy czym ostateczną decyzję może podejmować sam program, zwłaszcza
w procesach masowychi szybkozmiennych, lub posługujący się nim ekspert.

Podstawą uzyskania dobrej jakości wyników są jednak dobrej jakości dane. Wstępne przetwarzanie danych jest najbardziej kluczowym i często najtrudniejszym etapem eksploracji danych,
a jednocześnie etapem, któremu nie zawsze poświęca się należytej uwagi, co skutkuje niższą dokładnością całego systemu eksploracji danych. Dlatego prezentacja ta koncentruje się na przygotowaniu danych dobrej jakości poprzez selekcję informacji i redukcję szumu
z wykorzystaniem selekcji cech i wektorów.

Wstępne przetwarzanie danych jest najczęściej tym etapem eksploracji danych od którego najbardziej zależy jakość całego systemu. Bowiem jakość możliwych do uzyskania wyników jest ograniczona jakością samych danych i nawet najlepsza metoda nie da dobrych wyników, jeśli dane będą niskiej jakości, czyli będą zawierały dużo błędów pomiarowych, błędów przetwarzania, lub dużą ilość nie istotnych danych. To ostatnie dodatkowo wpływa na czas działania modeli, a czasami powoduje wręcz nie możliwość przeprowadzenia obliczeń ze względu na ograniczenia czasowe
i pamięciowe. Istotnym elementem wstępnego przetwarzania danych jest selekcja informacji celem poprawy dokładności przewidywania i redukcji rozmiaru danych.

Redukcję rozmiaru danych przeprowadza się w trzech podstawowych celach:

- aby umożliwić ich skuteczne dalsze przetwarzanie, które na pełnych danych może być niewykonalne ze względu na wymagany czas obliczeń.

- aby wybrać reprezentatywne dane, umożliwiające łatwą interpretację danych i procesów, które te dane opisują.

- nawet jeśli danych nie jest zbyt dużo, ani nie są zaszumiome, ich odpowiednia selekcja może poprawić działanie modelu.

Redukcję szumu natomiast przeprowadza się, aby wyeliminować błędne lub zniekształcone dane, które niekorzystnie wpłynęły by na działanie modelu.

Istnieją dwa podstawowe sposoby zarówno redukcji rozmiaru danych, jak i redukcji szumu w danych, którym poświęcona jest ta prezentacja:

- selekcja cech

- selekcja wektorów

Selekcja cech i wektorów może polegać na:

- pozostawieniu tylko najbardziej znaczących cech i wektorów.

- przypisaniu wag poszczególnym cechom i wektorom, tak, aby bardziej znaczące cechy i wektory
w większym stopniu wpływały na działanie końcowego modelu.

- wygenerowaniu na podstawie istniejących cech i wektorów nowego, mniejszego zbioru cech
i wektorów, który jednak zachowa większość właściwości oryginalnego zbioru.

Aby usprawnić proces predykcji, należy wykorzystać posiadaną wiedzę o procesie,
a w szczególności na podstawie istniejących cech skonstruować cechy, o których wiemy, że będą miały duże zdolności predykcyjne, np. jeśli interesują nas skutki zderzenia, a wiemy, lub przypuszczamy, że powinny być one proporcjonalne do energii kinetycznej, zaś oryginalnymy cechami są masa i prędkość obiektu, to należy skonstruować nową cechę wynoszącą m\*v\*v.

Ponadto dobrą praktyką, a nierzadko nawet koniecznością jest standardyzacja lub normalizacja wszystkich wielkości numerycznych. Cechy symboliczne należy dostosować do wymagań naszego algorytmu predykcji. Jeśli poszczególne wartości danej cechy tworzą logiczny ciąg, to można spróbować je zamienić na kolejne liczby np. całkowite w tej samej logicznej kolejności (np. mały = -1, średni = 0, duży = 1). Jeśli nasz algorytm predykcji nie obsługuje cech symbolicznych to należy je zamienić na numeryczne, albo jedna cecha -> jedna cecha numeryczna, albo jedna wartość cecha -> jedna cecha binarna.

**Demonstracja Programu RapidMiner**

Praktyczne demonstracje zostaną przeprowadzone w środowisku RapidMiner, który jest jednym
z najczęściej wykorzystywanych w przemyśle programów do predykcji i analizy danych (ponad 35.000 wdrożeń, w tym w firmach takich jak PayPal, Intel, Cisco, Ebay, Volkswagen, Deloitte, LuftHansa, Siemens). W środowisku RapidMiner cały proces można poskładać z bloków przy użyciu myszki. W najnowszej wersji 6.5, wersja podstawowa jest za darmo, natomiast wersje rozbudowane o obsługę większej ilości formatów baz danych i większej ilości pamięci są odpłatne. RapidMiner umożliwia załadowanie danych, wstępne przetwarzanie, kroswalidację, różne metody klasyfikacji i regresji, grupowania danych oraz inne operacje a także różne metody wizualizacji wyników i tworzenia raportów. Można rozszerzać jego funkcjonalność poprzez dodawanie nowych modułów dostępnych w internecie oraz poprzez samodzielne tworzenie modułów w języku java oraz poprzez jego integrację z własnym oprogramowaniem.



Teraz pokażemy, jak się posługiwać RapidMinerem. W tym celu utworzymy przykładowy proces, jak na powyższym rysunku i pokażemy również jeden z możliwych sposobów integracji RapidMinera z własnym programem napisanym w dowolnym języku. Główny proces wczytuje dane z pliku csv (Read CSV) i wykonuje kroswalidację (Validation). W ramach tej kroswalidacji jest część treningowa i testowa. W części treningowej najpierw dokonujemy selekcji wektorów algorytmem ENN (ENN selection) celem redukcji szumu, następnie algorytmem CA (CA selection) celem połączenia sąsiednich bardzo podobnych wektorów w jeden zawierający ich uśrednione wartości i zapisujemy zbiór wyselekcjonowanych wektorów do pliku csv (Write CSV). Następie uruchamiamy zewnętrzny program napisany w dowolnym języku (Execute Program), który wczytuje ten zbiór, przeprowadza na nim predykcję i zapisuje wynik do kolejnego pliku csv.
W części testowej RapidMiner wczytuje ten plik (Read CSV 2) i oblicza poprawność klasyfikacji (Performance), która jest następnie w głównym procesie logowana (Log) i uwzględniana przy obliczeniach wypadkowej dokładności w kroswalidacji. Następnie pokażemy integrację RapidMinera z własnym programem napisanym w Javie poprzez wywołanie w nim funkcji RapidMinera.

**Selekcja Cech – Filtry**

Istnieją trzy zasadnicze podejścia do selekcji cech: filtry, wrappery i metody wbudowane. Filtry wykorzystują miary statystyczne celem przypisania wag poszczególnym cechom. Pozwala to na odrzucenie cech o najniższych wagach. Filtry mogą być oparte m. in. o:

- współczynnik korelacji między poszczególnymi cechami a wielkością wyjściową

- Chi squared test

- information gain

Ranking cech utworzony przy pomocy filtra nie zawsze jest optymalny, ponieważ nie został zweryfikowany w praktyce, jednak jego zaletą jest prostota i szybkość działania. Jest to szczególnie istotne przy olbrzymich zbiorach danych, na których najpierw trzeba przeprowadzić redukcję wymiarowości, żeby się w ogóle dało przeprowadzić ich dalszą analizę.

Po wykonaniu rankingu często można również usunąć cechy, które są mocno skorelowane z innymi cechami. Choć w wielu przypadkach dobrze go działa, to jednak nie zawsze. Przykładem jest poniższy rysunek po lewej stronie, gdzie cecha x1 jest 100% skorelowana z cechą x2, ale najlepszą predykcję (klasa czerwona lub zielona) daje zastosowanie sumy obydwu tych cech x1+x2 (linia podziału jest nachylona pod kątem 45o do każdej z osi).



Dodatkowo kryteria oparte na współczynnikach korelacji mają taką wadę, że wykrywają tylko liniowe zależności między danym wejściem a wyjściem. Metodą na obejście tego problemu jest zastosowanie nieliniowego dopasowania wejścia do wyjścia poprzez różne przekształcenia wejścia, np. podniesienie do kwadratu, pierwiastek, logarytm, aproksymację odcinkową, itd. i następnie obliczenie współczynnika korelacji między tak przekształconą cechą wejściową a wyjściem.

Inną wadą metod rankingowych jest to, że nie uwzględniają współzależności między cechami. Czasami cechy, które samodzielnie mają niską korelację z wyjściem mogą znacząco poprawiać predykcję jeśli zostaną dodane do innej cechy. Przykład na powyższym rysunku po prawej stronie.

Powyższe rysunki pokazują pewne całkiem często występujące zjawisko: generując nowe cechy przez liniową kombinację istniejących cech (np x3 = a\*x1 + b\*x2, na rysunku 1: a=1 i b=1). można skutecznie zmniejszyć wymiarowość zbioru danych. Jedną z najpowszechniej stosowanych metod jest tu analiza składowych głównych – PCA (Principal Component Analysis), która z istniejących cech generuje nowe cechy i szereguje je według ilości informacji która zawiera dana cecha. Następnie pozostawia się jedynie najbardziej informatywne cechy, odrzucają resztę. Często działa to nadzwyczaj dobrze. Wadą tego podejścia jest natomiast to, że nowe cechy są trudne do interpretacji logicznej, np. co to znaczy x = 0.75\*napięcie + 0.34\*temperatura – 0.25\*ciśnienie? Więc jeśli potrzebujemy mieć możliwość prostego rozumienia takiego systemy i prostego wyciągania z niego reguł logicznych, to może to wykluczać taki sposób redukcji cech. Praktyczna demonstracja filtrów cech zostanie przeprowadzona w RapidMinerze.

**Selekcja Cech – Wrappery**

Aby zaradzić z wymienionymi niedoskonałością filtrów można zastosować wrappery, choć mają one wyższy nakład obliczeniowy. Wrappery działają razem z algorytmem uczenia. Proponowany jest pewien podzbiór zbioru cech i na tym podzbiorze cech uczony jest algorytm predykcyjny na zbiorze treningowym, po czym na zbiorze testowym (albo lepiej w kroswalidacji) sprawdzana jest dokładność predykcji. Następnie na podstawie tych wyników próbowany jest kolejny podzbiór cech i znowu sprawdzana dokładność. Jest to zadanie optymalizacji z dwoma kryteriami: maksymalizacja jakości predykcji (klasyfikacji, regresji, grupowania) i minimalizacja ilości cech. Każdemu z tych kryteriów przypisujemy pewną wagę. W granicznym przypadku może być zerowa waga przypisana do ilości cech. W ten sposób selekcja cech będzie jedynie dokonana celem zwiększenia dokładności modelu. Ilość możliwych kombinacji cech wynosi

$$K= \sum\_{n=1}^{F}\left(\frac{F}{n}\right)=\sum\_{n=1}^{F}\frac{F!}{n!\left(F-n\right)!}$$

gdzie F jest liczbą cech, przykładowo dla F=10: K= 1.0E3, dla F=30:K=1.1E9 dla F=70: K=1.2E21, dla F=170: K=1.5E51. Przy ilości cech powyżej kilkunastu przeszukanie całej przestrzeni rozwiązań jest praktycznie nie wykonalne. Dlatego stosuje się różne metody przeszukiwań, jak best-first search, beam-search, random hill-climbing, a nawet algorytmy genetyczne.

Przy lokalnych metodach poszukiwań można stosować przeszukiwanie w przód (forward selection) lub wstecz (backward selection). W pierwszym przypadku zaczynamy od pojedynczej cechy
i dokładamy kolejne cechy, które najbardziej poprawiają jakość modelu. W drugim przypadku zaczynamy od pełnego oryginalnego zbioru cech i usuwamy kolejne cechy, których usunięcie najbardziej poprawi wynik. Jeśli zależy nam bardziej na redukcji cech, to możemy usuwać kolejne cechy nawet gdy ich usuwanie powoduje już pogorszenie wyniku, ale usuwamy te, których usunięcie najmniej pogarsza wynik.

Często selekcja w przód daje lepsze rezultaty: większą dokładność predykcji przy mniejszej ilości cech. Ale nie zawsze, gdyż w niektórych przypadkach nie znajduje ona najlepszej konfiguracji jeśli ta jest wyrażona poprzez kilka cech, z których każda samodzielnie ma słabe właściwości predykcyjne, ale całość jako grupa ma bardzo dobre, jak na rysunkach na poprzedniej stronie. Praktyczna demonstracja wrapperów do selekcji cech zostanie przeprowadzona
w RapidMinerze.

**Selekcja Cech – Metody wbudowane**

W metodach wbudowanych selekcja cech jest integralną częścią algorytmu predykcyjnego. Zatem jednocześnie jest uczony model predykcyjny i wykonywana selekcja cech. Pod względem efektywności obliczeniowej jest to bardzo dobre rozwiązanie. Jeżeli jednak zbiór danych jest bardzo duży może to być czasowo niewykonalne i wówczas trzeba użyć filtrów cech. Np. drzewa decyzyjne wykonują forward selection. Regresja liniowa sama w sobie nie wynonuje selecji cech, ale można odrzucić te cechy, dla których przyjmnie najmniejsze wagi. Podobnie przy sieci neuronowej – można odrzucić te cechy, dla których suma wag wszystkich neuronów pierwszej warstwy ukrytej będzie najmniejsza. Choć w przypadku sieci neuronowej, dokładność tego rozwiązani najczęściej ustępuje wrapperom, czy nawet filtrom cech. Dodatkowo w metodach wbudowanych można używać regularyzacji poprzez stosowanie dodatkowego członu kary za duży model w funkcji celu lub bardziej złożonych algorytmów regularyzacyjnych. Demonstracja metod wbudowanych do selekcji cech zostanie przeprowadzona w RapidMinerze.

**Selekcja wektorów w zagadnieniach klasyfikacji**

W przypadku selekcji wektorów, w przeciwieństwie do selekcji cech nie stosuje się na ogół jednego wspólnego podejścia do redukcji szumu i zmniejszenia rozmiaru danych, co więcej stosuje się różne metody dla klasyfikacji i regresji.

W zagadnieniu klasyfikacji istotne jest odpowiednie wyznaczenie granic decyzjnych między klasami. Do tego potrzebne są tylko wektory położone w pobliżu granicy klas zaznaczonej ukośną czarną linią na poniższym rysunku po lewej stronie. Linia ta oddziela klasę zieloną od czerwonej. Wektory położone w szarych kołach nie są potrzebne do wyznaczenia tej granicy i można je usunąć. Natomiast wektory będące w małych żółtych kółkach należą do innej klasy, niż wszystkie otaczające je wektory, są to jakieś błędne dane, które stanowią szum. Także i te dane należy usunąć. W pierwszym przypadku usuwaliśmy wektory celem zmniejszenia rozmiaru zbioru. W drugim przypadku natomiast celem redukcji szumu.



Najprostszą metodą zmniejszenia rozmiaru zbioru jest CNN (Condensed Nearest Neighbor). Polega ona na tym, że wektor, którego rzeczywista klasa jest taka sama, jak klasa większości jego najbliższych sąsiadów zostaje zaznaczony do usunięcia. Po przejściu iteracji przez wszystkie wektory wektory, które zostały oznaczone do usunięcia się ostatecznie usuwa.

Z kolei najprostszą metodą redukcji szumu, czyli usuwania wektorów zbytnio różniących się od swoich sąsiadów jest ENN (Editted Nearest Neighbour). Analogicznie jak w poprzednim wypadku przewidujemy tu klasę danego wektora. Natomiast teraz zaznaczamy do usunięcia te wektory których rzeczywista klasa jest inna, niż klasa przewidziana i usuwamy je po zakończeniu iteracji.

Do przewidzenia klasy danego wektora w klasycznych algorytmach CNN i ENN używa się algorytmu k-NN. Przy czym podstawowy k-NN wyznacza odległość między dwoma wektorami, traktując wszystkie cechy równorzędnie. A zatem pierwsza bezwzględnie wymagana rzecz, to standardyzacja, lub normalizacja danych. Druga rzecz, która najczęściej mocno poprawia wyniki, to dokonanie selekcji cech przed użyciem k-NN. Użycie filtra opartego na korelacji i następnie użycie ważonej odległości, gdzie waga składowej od każdej cechy jest proporcjonalna do jej korelacji z wyjściem znacząco poprawia działanie k-NN, a zatem poprawia i wynik selekcji wektorów, gdyż usuwane są bardziej odpowiednie wektory.

Oczywiście istnieją bardziej zaawansowane metody selekcji wektorów dające większą kontrolę nad prowadzonym procesem. Wspomnimy o wybranych z nich. Praktyczna demonstracja selekcji wektorów dla zagadnienia klasyfikacji zostanie przeprowadzona w RapidMinerze.

**Selekcja wektorów w zagadnieniach regresji**

Na prawym rysunku na poprzedniej stronie wysokość słupka odpowiada wartości wyjścia danego wektora. Wektory zaznaczone na czerwono nie pasują do swoich sąsiadów, jeden z nich ma znacznie większą wartość wyjścia, a drugi znacznie mniejszą. Należy je więc usunąć ze zbioru, gdyż są to błędne dane. Natomiast w przypadku redukcji wymiarowości można usunąć zawsze jeden z dwóch wektorów, które są zarówno bardzo blisko położone w przestrzeni wejść, jak i mają bardzo zbliżoną wartość wyjściową (zaznaczone zieloną elipsą). W przypadku regresji tylko niewiele wektorów jesteśmy w stanie usunąć eliminując podobne wektory, gdyż wyznaczamy tu wartość wyjść dla każdego punktu w przestrzeni wejść, w przeciwieństwie do klasyfikacji, gdzie można było często usunąć wszystkie wektory położone z dala od granicy klas.

Opracowano ostatnio adaptacje algorytmu CNN do zagadnień regresji. Istnieją tutaj dwie możliwości do wyboru: albo używamy jakiegoś progu odległości i jeżeli wartość przepowiedziana przez k-NN (lub inny algorytm predykcji) różni się nie więcej niż o ten próg od wartości rzeczywistej to możemy taki wektor zaznaczyć do usunięcia. W praktyce wartość progu można ustawić na ok. 0.1-0.2 standardowych odchyleń z wyjść k najbliższych sąsiadów tego punktu. Drugą opcją jest przeprowadzenie dyskretyzacji wyjść i zamienienie tego zagadnienia na zagadnienie klasyfikacji wieloklasowej. Aczkolwiek, ze względu na problemy na granicy klas, pierwsze podejście często działa lepiej. Po przejściu iteracji przez wszystkie wektory wektory, które zostały oznaczone do usunięci się ostatecznie usuwa.

Stosując analogiczną adaptację algorytmu ENN do regresji celem redukcji szumu, również przewidujemy jego wartość wyjściową danego wektora. Natomiast teraz zaznaczamy do usunięcia te wektory których rzeczywista wartość wyjścia różni się od wartości przewidzianej więcej niż o zadany próg i usuwamy je po zakończeniu iteracji. W praktycznych zastosowaniach dobrze sprawdza się wartość progu równa ok. 5-8 standardowych odchyleń z wyjść k najbliższych sąsiadów tego punktu.

I tu, podobnie, jak przy klasyfikacji można stosować bardziej zaawansowane metody, niż ENN
i CNN zaadoptowany do regresji. Wspomnimy o nich w prezentacji. Ponadto wybrane wektory nie koniecznie trzeba odrzucać, tylko można im przypisywać różną wagę, używaną potem przez algorytm precykcji. Praktyczna demonstracja selekcji wektorów dla zagadnienia regresji zostanie teraz przeprowadzona w RapidMinerze.

Mamy tu zagadnie optymalizacji dwu-kryterialnej. Jednym kryterium jest maksymalizacja jakości predykcji, a drugim minimalizacja ilości wektorów. Często początkowemu zmniejszeniu ilości wektorów towarzyszy wzrost dokładności predykcji, lecz ich dalsza redukcja powoduje spadek dokładności. W zależności od istotności rozmiaru zbioru względem jakości predykcji możemy usuwać więcej, lub mniej wektorów (choćby przez dobór omawianego progu).

Można stosować podejście podobne do wrapperów w selekcji cech, ale nie bazując na pojedynczym wektorze, lecz zmniejszając kolejno ilość wektorów poprzez coraz ostrzejsze progi selekcji. Można również zastosować metody wbudowane. Np. sieć neuronowa ma taką właściwość, że popełnia największe błędy na wektorach najbardziej odstających, a zwłaszcza, gdy zamiast standardowej miary błędu MSE zastosuje się specjalne miary błędu odporne na szum, można te wektory stanowiące szum wyłapać w zagadnieniach regresji i to akurat działa bardzo dobrze. W przypadku klasyfikacji sieć taka generuje natomiast najmniejszy błąd dla wektorów będących z dala od granic decyzyjnych, w ten sposób można je zlokalizować i usunąć, choć to nie zawsze już działa tak dobrze.

**Integracja selekcji cech z selekcją wektorów**

Selekcja cech i selekcja wektorów są zagadnieniami wzajemnie powiązanymi ponieważ odrzucenie pewnych cech może zmieniać zbiór wektorów przeznaczony do odrzucenia, jak i odrzucenie pewnych wektorów może zmieniać zbiór cech które należy odrzucić (jak to było już zasygnalizowane z k-NN z ważonymi cechami).

Zostało to już podjęte pewne prace zmierzające do integracji selekcji cech z selekcją wektorów. Początkowe prace proponowały różne algorytmy genetyczne, gdzie w jednej części chromosomu były zakodowane cechy, a w innej wektory. Natomiast funkcją kryterialną była suma jakości predykcji (poprawność klasyfikacji lub RMSE dla regresji na zbiorze testowym), ilości pozostających cech przemnożonych przez pewną wagę, ilości pozostawionych wektorów przemnożonych przez inną wagę. Dobierając odpowiednie wagi można bardziej zwracać uwagę na jakość rozwiązania lub na mały rozmiar zbioru. Dla małych zbiorów podejście takie może skutkować znalezieniem bardzo dobrych rozwiązań. Souza et. al zaproponowali także niezależną selekcję wektorów i cech przy pomocy metod ewolucyjnych i po każej iteracji ocenie gotowego rozwiązania. Zmniejszyło to czas obliczeń, ale i tak przedstawili wyniki tylko dla bardzo małych zbiorów. Np. dla zbioru Ionosphere (351 wektorów, 34 cechy) czas obliczeń wynosił ok. 3 godzin na komputerze z 2008 roku.

Oczywistą bowiem wadą takiego podejścia jest bardzo wysoki koszt obliczeniowy, co wyklucza go
z praktycznych zastosowań dla zagadnień z setkami tysięcy wektorów, zwłaszcza, że złożoność obliczeniowa rośnie szybciej, niż liniowo, bo dłuższy chromosom wymaga użycia większej ilości osobników w populacji i przez to z reguły większej ilosci iteracji. Nawet w najbardziej optymistycznym scenariuszu przyjmując złożoność samego algorytmu ewolucyjnego O(n2)
i modelu predykcyjnego O(n) dostajemy łącznie złożoność w najlepszym razie O(n3).

Dlatego można wykonywać iteracyjnie na przemian coraz mocniejszą selekcję cech i wektorów. Tj. stosując backward selection usunąć jedną cechę, następnie usunąć kilka wektorów stosując najpierw ENN potem CNN ale z kryteriami dla klasyfikacji tak, że np. z 9-NN musi być 7-NN by zadziałało, podobnie z kryteriami dla regresji ustawić wysokie progi. Następna iteracja znowu usuwa jedną cechę w backward elimination i powtarza selekcję wektorów z ostrzejszymi progami. Zaletą takiego podejścia jest to, że operacja na zbiorze wielkości Ionosphere trwa na pojedynczym procesorze kilkadziesiąt sekund.

Odpowiednio zmodyfikowane sieci neuronowe mogą jednocześnie dokonywać selekcji cech
i wektorów, a zwłaszcza w zagadnieniach regresji. Przy czym trzeba to zaznaczyć, że nie w tym celu sieci neuronowe zostały skonstruowane i nawet ich modyfikacje nie pozwalają na ogół na osiągnięcie tak dobrych rezultatów, jak metody dedykowane do selekcji cech i wektorów.. Ponadto jest to możliwe tylko w przypadku, gdy rozmiar zbioru nie uniemożliwia uczenia na nim sieci neuronowej. Jednakże zaletą użycia sieci neuronowej jest tutaj prostota budowy systemu: rezygnujemy z dedykowanego modułu selekcji cech i wektorów. Selekcję wektorów wykonujemy poprzez odrzucenie tych, na których nauczona sieć robi największy błąd, a w klasyfikacji także tych, na których robi najmniejszy błąd, a selekcję cech poprzez eliminacje tych cech, a których są najmniejsze wagi w pierwszej warstwie ukrytej. Podobnie drzewa decyzyjne można użyć do selekcji cech poprzez odrzucenie tych cech, na których nie zostają nigdy dokonane punkty podziału, lub przynajmniej nie zostają dokonanew takich liściach, które zawierają sumarycznie daną ilość wektorów dla danej cechy i odrzucenie tych wektorów, które są błędnie sklasyfikowane, lub na których jest największy błąd w przypadku drzew regresyjnych.