ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 733 STUDIA INFORMATICA NR 30 2012

PAWEŁ ZIEMBA[*]
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

REDUKCJA WYMIAROWOŚCI I SELEKCJA CECH W ZADANIACH KLASYFIKACJI I REGRESJI Z WYKORZYSTANIEM UCZENIA MASZYNOWEGO

Wprowadzenie

Uczenie maszynowe wykorzystywane jest w zadaniach regresji i klasyfikacji. W zadaniach uczenia maszynowego na podstawie określonych cech obiektu przewidywana jest jego wartość ogólna lub przynależność do określonej klasy. Jest to wykonywane w oparciu o wcześniej przeprowadzony proces trenowania, w trakcie którego algorytm klasyfikacyjny „uczy się", jakie są rzeczywiste klasy obiektów treningowych i powiązuje przynależność do określonej klasy z wartościami cech obiektów. Jednym z podstawowych problemów w zadaniach klasyfikacji jest wielowymiarowość obiektu, który ma zostać przypisany do określonej klasy lub dla którego przypisywana jest wartość. Wymiarowość stanowi poważną przeszkodę dla efektywności algorytmów eksploracji danych i uczenia maszynowego. Problem ten nosi miano *przekleństwa wymiarowości*[1]. Redukcja wymiarów obiektów poddawanych klasyfikacji pozwala na: poprawę wyników predykcji, zmniejszenie wyma-

[*] pziemba@wi.ps.pl.

[1] B. Chizi, O. Maimon, *Dimension Reduction and Feature Selection*, w: *Data Mining and Knowledge Discovery Handbook*, red. O. Maimon, L. Rokach, Springer, Nowy Jork 2010, s. 83–100.

gań obliczeniowych, zmniejszenie wymagań odnośnie gromadzenia danych, redukcję kosztów przyszłych pomiarów, poprawę jakości danych[2].

Redukcja wymiarów zadania klasyfikacyjnego może być wykonywana przez proces selekcji cech, który koncentruje się na określeniu pewnych cech w zbiorze danych jako istotnych i odrzuceniu nadmiarowych zmiennych. Wybierany jest podzbiór charakterystyk z oryginalnego zbioru cech. W tym celu wykorzystywane są różnorodne algorytmy, oceniające poszczególne cechy względem określonego kryterium opisującego ich znaczenie w zadaniu klasyfikacji[3].

W artykule szczegółowo omówiono zagadnienia związane z wykorzystaniem metod selekcji cech do redukcji wymiarowości. W kolejnych rozdziałach przedstawiono: podstawowe założenia selekcji cech, podział metod wykorzystujących ten proces, wybrane algorytmy selekcji oraz przykład obliczeniowy.

1. Podstawowe zagadnienia związane z selekcją cech

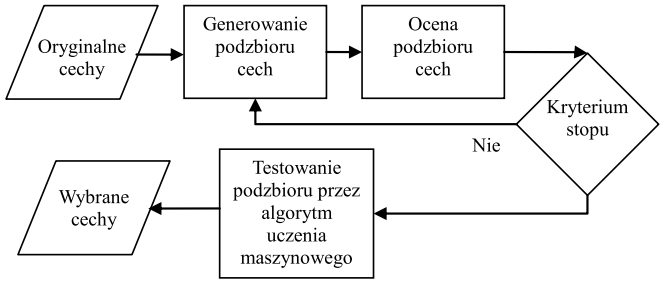
Proces selekcji cech można traktować jako problem przeszukiwania zbioru charakterystyk opisujących obiekt poddawany klasyfikacji według pewnego kryterium oceny. Metody selekcji cech złożone są zazwyczaj z czterech elementów (kroków), takich jak: generowanie podzbioru cech, ocena podzbioru, kryterium stopu, walidacja rezultatów[4]. Metody selekcji cech dzielą się na dwa rodzaje procedur, to jest: filtry i wrappery. Między tymi grupami metod występują istotne różnice. Filtry bazują na niezależnej ocenie cech z wykorzystaniem ogólnych charakterystyk danych. Mogą być tu wykorzystywane na przykład współczynniki korelacji między wartościami cech a przynależnością do określonej klasy. Zbiór cech obiektu jest poddawany filtracji w celu określenia najbardziej obiecującego podzbioru atrybutów przed rozpoczęciem trenowania algorytmu uczenia maszynowego[5].

[2] I. Guyon, *Practical Feature Selection: from Correlation to Causality*, w: *Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security*, red. F. Fogelman-Soulié, D. Perrotta, J. Piskorski, R. Steinberger, IOS Press, Amsterdam 2008, s. 27–43.

[3] D. Hand, H. Mannila, D. Smyth, *Eksploracja danych*, WNT, Warszawa 2005, s. 414–416.

[4] H. Liu, L. Yu, H. Motoda, *Feature Extraction, Selection, and Construction*, w: *The Handbook of Data Minin*g, red. N. Ye, Lawrence Erlbaum Associates, Mahwah 2003, s. 409–424.

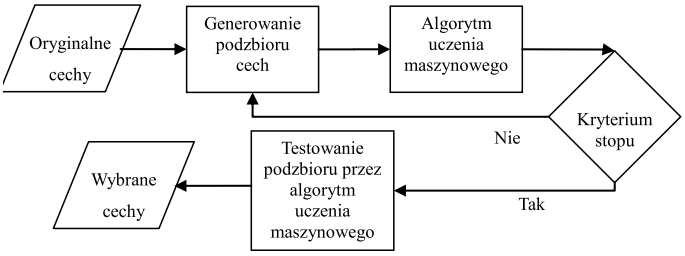
[5] I.H. Witten, E. Frank, *Data Mining. Practical Machine Learning Tools and Techniques*, Elsevier, San Francisco 2005, s. 288–295.



Rys. 1. Struktura algorytmów należących do kategorii filtrów

Źródło: M.A. Hall, G. Holmes, *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*, „IEEE Transactions on Knowledge and Data Engineering" 2003, nr 3, s. 1437–1447.

Wrappery z kolei oceniają poszczególne podzbiory cech z wykorzystaniem algorytmów uczenia maszynowego, które ostatecznie zostaną wykorzystane w zadaniu klasyfikacji bądź regresji. Algorytm uczący jest w tym przypadku zawarty w procedurze selekcji cech, a do oszacowania dokładności klasyfikatora korzystającego z określonego podzbioru cech wykorzystywana jest zazwyczaj walidacja krzyżowa. Struktury algorytmów należących do kategorii filtrów i wrapperów przedstawiono na rysunkach 1 i 2.



Rys. 2. Struktura algorytmów należących do kategorii wrapperów

Źródło: jak pod rys. 1.

Generowanie podzbioru atrybutów może odbywać się na różne sposoby. Najbardziej podstawowymi strategiami są: tworzenie indywidualnego rankingu, przeszukiwanie w przód i przeszukiwanie wstecz. Tworzenie indywidualnego rankingu rozpoczyna się przy pustym podzbiorze cech. W każdym kroku do tego podzbioru dodawany jest atrybut określony jako najlepszy bez uwzględnienia ewentualnych zależności między cechami. Od pustego podzbioru atrybutów rozpoczyna się również procedura przeszukiwania w przód. W pierwszym kroku do podzbioru dodawana jest cecha uznana za najlepszą bez uwzględnienia zależności między cechami. W kolejnym kroku do podzbioru atrybutów dodawany jest atrybut, który wraz z wybranym wcześniej tworzy najlepszą parę cech. Procedura ta przebiega iteracyjnie, aż do osiągnięcia kryterium stopu. Procedura przeszukiwania wstecz rozpoczyna się, gdy podzbiór jest kopią pierwotnego zbioru cech. W tym przypadku kolejno z podzbioru usuwane są cechy, których usunięcie skutkuje maksymalizacją kryterium oceny podzbioru. Procedura indywidualnego rankingu nie uwzględnia zależności między atrybutami, lecz rozpatruje oddzielnie każdą z cech obiektu. W związku z tym, może ona dawać gorsze rezultaty od pozostałych omówionych strategii. W procedurze przeszukiwania wstecz brane jest pod uwagę więcej możliwych współzależności pomiędzy cechami, jednak jest ona bardziej złożona obliczeniowo od strategii przeszukiwania w przód[6].

Wrappery różnią się między sobą tylko zastosowanymi algorytmami uczenia maszynowego, więc rezultaty uzyskane z ich wykorzystaniem zależą wyłącznie od jakości algorytmu uczenia maszynowego i dopasowania algorytmu do określonego zadania klasyfikacyjnego. Bardziej interesujące wydają się procedury filtracyjne, określające istotność poszczególnych atrybutów za pomocą innych miar niż stopień poprawnych klasyfikacji, w związku z czym w artykule zbadano działanie algorytmów generowania i oceny podzbiorów cech, stosowanych w procedurach filtracyjnych.

[6] K. Michalak, H. Kwaśnicka, *Correlation-based feature selection strategy in classification problems*, „International Journal of Applied Mathematics and Computer Science" 2006, nr 4, s. 503–511.

2. Metody filtracyjne stosowane do selekcji cech

Wśród metod wykorzystujących do selekcji cech filtry, stosowane są między innymi procedury: ReliefF, LVF, FCBF, CFS, istotności atrybutów (ang. *significance attribute*).

Podstawową ideą metody ReliefF jest ocena atrybutów według tego, jak dobrze pozwalają one rozróżnić podobne obiekty, to znaczy takie, które znajdują się blisko siebie z perspektywy podobieństwa wartości cech. Stosowana jest tutaj metoda najbliższych sąsiadów, a więc również funkcja odległości. W procedurze ReliefF wykorzystywana jest heurystyka mówiąca o tym, że dobry atrybut powinien rozróżniać leżące blisko siebie obiekty należące do innych klas, a dodatkowo powinien mieć taką samą wartość dla leżących blisko siebie obiektów należących do tej samej klasy. Dla każdego obiektu *r* i każdej analizowanej cechy X_i znajdowanych jest *k* obiektów $s_{1...k}$ tej samej klasy co obiekt *r*, dla których analizowana cecha X_i jest najbardziej zbliżona do tej samej cechy X_i w badanym obiekcie *r*. Dodatkowo dla każdego obiektu *r* i każdej analizowanej cechy X_i znajdowanych jest *k* obiektów $s_{1...k}$, należących do innej klasy niż obiekt *r*, dla których wartość cechy X_i jest najbliższa wartości cechy X_i w obiekcie *r*. W tym przypadku znajdowanych jest łącznie *k*(c-1)* obiektów (gdzie *c* oznacza liczbę rozpatrywanych klas) lub, inaczej mówiąc, znajdowanych jest *k* najbliższych obiektów odrębnie w każdej z klas, do których nie należy obiekt *r*. Metodę wyznaczania oceny cechy X_i można opisać wzorem (1) [7]:

$$q(X_i) = \sum_r \sum_{C(s)\neq C(r)} \left(\frac{P(C(s))}{1-P(C(r))} * \frac{d_{rs,i}}{k} \right) - \sum_r \sum_{C(s)=C(r)} \frac{d_{rs,i}}{k}.$$

Funkcja odległości opisana jest wzorem (2)[8]:

$$d_{rs,i} = \begin{cases} 0 & gdy\ r(i) = s(i) \\ 1 & gdy\ r(i) \neq s(i) \\ \dfrac{r(i)-s(i)}{nu(i)} & \end{cases} dla\ atrybutów\ \ nominalnych.$$

[7] I. Kononenko, S.J. Hong, *Attribute Selection for Modeling*, „Future Generation Computer Systems" 1997, nr 2–3, 1997, s. 181–195.

[8] K. Kira, L.A. Rendell, *A Practical Approach to Feature Selection*, ML92 Proc. of IWML 1992, s. 249–256.

Metoda LVF wykorzystuje podejście probabilistyczne w celu określenia kierunku wskazującego prawidłowe rozwiązanie. Do prowadzenia wyszukiwania rozwiązania wykorzystywana jest tutaj losowość gwarantująca uzyskanie akceptowalnego rozwiązania nawet w sytuacji, gdy w trakcie poszukiwania najlepszego podzbioru podejmowane są błędne decyzje[9]. Algorytm LVF na wstępie generuje losowy podzbiór cech, a następnie określa jego spójność przez znalezienie minimalnej liczby cech, które różnicują przynależność obiektów do poszczególnych klas w taki sam sposób, jak pełny zestaw cech. Wykorzystywane jest tutaj kryterium niespójności, które określa stopień akceptacji danych o zredukowanej wymiarowości. Współczynnik niespójności danych opisanych z wykorzystaniem zredukowanego podzbioru cech jest w tym przypadku porównywany z współczynnikiem niespójności danych charakteryzowanych przez pełny zbiór cech. Zredukowany podzbiór cech jest akceptowany, gdy uzyskany dla niego współczynnik niespójności jest niższy od tego samego współczynnika określonego dla pełnego zbioru atrybutów. Współczynnik niespójności można określić za pomocą wzoru (3):

$$IncR = \frac{\sum_i D_i - M_i}{N},$$

gdzie:

D_i jest liczbą wystąpień i-tej kombinacji wartości cech,

M_i określa liczność obiektów klasy dominującej dla *i*-tej kombinacji atrybutów,

N określa liczbę obiektów.

Procedura FCBF bazuje na współczynnikach korelacji, a dokładniej mówiąc, na symetrycznej niepewności (ang. *symmetrical uncertainty*). Symetryczna niepewność określona jest jako stosunek zawartości informacyjnej pary atrybutów do sumy entropii tych atrybutów i opisuje ją wzór (4):

$$SU(X,Y) = 2 * \left[\frac{IG(X \mid Y)}{H(X) + H(Y)} \right].$$

[9] H. Liu, R. Setiono, *A Probabilistic Approach to Feature Selection – A Filter Solution*, Proc. of ICML 1996, s. 319–327.

Zawartość informacyjna mówi o tym, jaki zysk informacji uzyskuje się, wykorzystując entropię cechy X w porównaniu do entropii cechy X po obserwacji cechy Y. Entropia z kolei określa niepewność zmiennej losowej[10]. Dodatkowo w metodzie FCBF stosowane są pomocniczo, oddzielnie dla każdej cechy, zbiory cech nadmiarowych. Zbiór Sp_i^+ zawiera cechy nadmiarowe względem cechy F_i, mające wyższy od F_i współczynnik symetrycznej niepewności w powiązaniu z klasą *C*. Z kolei zbiór Sp_i^- zawiera cechy nadmiarowe względem F_i, mające niższy od niej współczynnik symetrycznej niepewności w powiązaniu z klasą *C*. Na wstępie procedury FCBF obliczane są symetryczne niepewności dla każdej cechy, a do dalszego rozpatrzenia wybierane są tylko atrybuty, których symetryczne niepewności są większe od przyjętego progu. Umieszczane są one w zbiorze S' w porządku malejącym, opartym na wartościach ich symetrycznych niepewności. Następnie zbiór S' jest sprawdzany pod względem występowania w nim nadmiarowości cech. Ewentualne nadmiarowe cechy są z niego usuwane z wykorzystaniem trzech heurystyk:

1. Jeżeli zbiór Sp_i^+ jest pusty, cecha F_i uznawana jest za cechę główną. Przerywana jest identyfikacja cech nadmiarowych względem cech zawartych w zbiorze Sp_i^- i zawartość zbioru Sp_i^- jest usuwana.
2. Jeżeli zbiór Sp_i^+ nie jest pusty, przed podjęciem decyzji odnośnie cechy F_i, przetwarzany jest zbiór wszystkich cech w Sp_i^+. Jeżeli żadna cecha znajdująca się w zbiorze Sp_i^+ nie jest uznawana za główną, wykonywana jest heurystyka 1. W przeciwnym przypadku cecha F_i jest usuwana z S' i na podstawie pozostałych cech zawartych w S' podejmowana jest decyzja, czy usunąć zawartość zbioru Sp_i^-.
3. Cecha z największą wartością symetrycznej niepewności jest zawsze cechą główną i punktem startowym procedury usuwania nadmiarowych cech[11].

Procedura ta kończy się po wykonaniu jej dla każdego atrybutu zawartego w zbiorze S'.

Metoda CFS, podobnie jak FCBF, oparta jest na badaniu korelacji pomiędzy cechami. Globalną miarą korelacji wykorzystywaną w procedurze CFS

[10] S.S. Kannan, N. Ramaraj, *A novel hybrid feature selection via Symmetrical Uncerteinty ranking based local memetic search algorithm*, „Knowledge-Based Systems" 2010, nr 23, s. 580–585.

[11] L. Yu, H. Liu, *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, Proc. of ICML 2003, s. 856–863.

jest korelacja liniowa Pearsona, natomiast lokalnie stosowana jest symetryczna niepewność. Wykorzystywana jest tutaj heurystyka mówiąca o tym, że dobry podzbiór cech zawiera atrybuty silnie skorelowane z określoną klasą obiektów oraz nieskorelowane z innymi klasami i atrybutami. Pozwala ona odfiltrować cechy, które w niewielkim stopniu opisują przynależność obiektu do określonej klasy oraz cechy nadmiarowe, silnie powiązane z innymi cechami. W metodzie CFS stosowany jest wzór (5)[12]:

$$Merit_s = \frac{k * r_{cf}}{\sqrt{k + k * (k-1) * r_{ff}}},$$

gdzie:

$Merit_s$ jest wartością heurystyki dla podzbioru S zawierającego k cech,

r_{cf} jest średnią wartością współczynnika korelacji pomiędzy cechami podzbioru S i klasami obiektów,

r_{ff} określa średnią korelację wzajemną między cechami.

Licznik wzoru (4) określa, jak dobrze podzbiór cech pozwala przewidzieć przynależność obiektu do danej klasy, natomiast mianownik opisuje nadmiarowość w zbiorze cech. W procedurze CFS na wstępie, przez obliczenie symetrycznej niepewności, wyznaczana jest macierz korelacji wzajemnych między atrybutami oraz korelacji między atrybutami a klasami obiektów. Następnie wykonywane jest przeszukiwanie w przód za pomocą algorytmu *best first*. Przeszukiwanie kończy się, gdy kolejne pięć rozwinięć podzbioru cech nie przynosi poprawy rezultatów[13].

W metodzie istotności atrybutów wykorzystywane są współczynniki dwukierunkowych powiązań pomiędzy atrybutami i przynależnościami obiektów do klasy. Metoda ta oparta jest na heurystyce mówiącej o tym, że jeżeli atrybut jest istotny, to istnieje duże prawdopodobieństwo, że obiekty dopełniające zbiory wartości tego atrybutu będą należały do dopełnienia zbiorów klas. Dodatkowo, przy założeniu, że klasy decyzyjne dla dwóch zbiorów obiektów są różne, można oczekiwać, że wartości istotnych atrybutów

[12] M.A. Hall, L.A. Smith, *Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper*, Proc. of IFAIRSC 1999, s. 235–239.

[13] M.A. Hall, *Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning*, Proc. of ICML 2000, s. 359–366.

w obiektach należących do tych dwóch zbiorów będą różne. Istotność każdego z atrybutów jest wyznaczana jako średnia wartość ogólnych powiązań danego atrybutu z klasami (AE) oraz klas z danym atrybutem (CE). Atrybut jest istotny, gdy wartości obydwu powiązań są wysokie. Wartość powiązania atrybutu z klasami (AE) odzwierciedla łączny wpływ wszystkich możliwych wartości atrybutu i ich związki z określonymi klasami obiektów. Współczynnik powiązania klas z atrybutem (CE) określa, w jaki sposób zmienia się wartość atrybutu zależnie od zmiany klasy obiektu. Wartości AE i CE opisane są wzorami (6) i (7):

$$AE(X_i) = \left(\frac{1}{k} * \sum_{r=1}^{k} \left(P_i^r(w) + P_i^{\sim r}(\sim w) \right) \right) - 1,$$

gdzie:

k oznacza liczbę możliwych wartości i-tego atrybutu,

$P_i^r(w)$ i $P_i^{\sim r}(\sim w)$ kolejno oznaczają prawdopodobieństwo tego, że obiekty z określoną r-tą wartością atrybutu X_i należą do klas z określonego podzbioru klas w; prawdopodobieństwo tego, że obiekty z wartością atrybutu X_i różną od r-tej wartości nie należą od klas zawartych w podzbiorze klas w.

$$CE(X_i) = \left(\frac{1}{m} * \sum_{j=1}^{m} \left(P_i^j(V_i^j) + P_i^{\sim j}(V_i^{\sim j}) \right) \right) - 1,$$

gdzie:

m oznacza liczbę klas,

$P_i^j(V)$ określa prawdopodobieństwo tego, że wartości atrybutu A_i obiektów należących do klasy j zawarta jest w podzbiorze V,

$P_i^{\sim j}(V_i^{\sim j})$ opisuje prawdopodobieństwo tego, że obiekty nienależące do klasy j mają wartość atrybutu A_i, która nie jest zawarta z podzbiorze V[14].

[14] A. Ahmad, L. Dey, *A feature selection technique for classificatory analysis*, „Pattern Recognition Letters" 2005, nr 26, s. 43–56.

3. Procedura badawcza i wyniki badań

W przyjętej procedurze badawczej koncentrowano się na wygenerowaniu rankingów cech za pomocą każdej z omówionych metod. Badanie miało wskazać różnice w rankingach uzyskiwanych z wykorzystaniem każdej procedury. Wykorzystano trzy zbiory danych pochodzące z serwisu UCI Machine Learning Repository[15]: Car Evaluation Data Set, Image Segmentation Data Set oraz Wine Quality Data Set[16]. Zbiór Car Evaluation zawiera 1728 obiektów, z których każdy jest opisany za pomocą 6 atrybutów o dyskretnych wartościach i może należeć do jednej z 4 klas określających dopuszczalność zakupu samochodu, przy czym w każdej klasie jest różna liczba obiektów. Zbiór Image Segmentation zawiera 2100 obiektów opisanych z wykorzystaniem 19 atrybutów. Każdy obiekt może należeć do jednej z 7 klas określających zawartość obrazu graficznego opisywanego przez obiekt. W każdej z klas zawarta jest taka sama liczba obiektów, a wykorzystane atrybuty mają wartości ciągłe. Jeżeli chodzi o zbiór Wine Quality, wykorzystano tylko jego część, która opisuje przynależność win białych do jednej z 10 klas jakościowych. Stosowanych jest tutaj 11 atrybutów ciągłych, a wykorzystana część zbioru zawiera 4898 obiektów. Jeżeli chodzi o zastosowane strategie generowania podzbioru atrybutów, metody LVF i CFS korzystały z przeszukiwania w przód, natomiast pozostałe trzy metody stosowały strategię rankingu indywidualnego. Dla metody ReliefF do oceny atrybutu zastosowano 10 najbliższych sąsiadów, a próbkowanie było wykonywane na wszystkich obiektach. Rankingi istotności cech uzyskane za pomocą każdej z omówionych metod dla poszczególnych zbiorów danych zawarte są w tabelach 1, 2 i 3.

Kolejności w rankingach cech dla zbioru Car Evaluation są takie same z użyciem każdej metody z wyjątkiem procedury LVF. Metoda ta daje ranking różniący się trzema pierwszymi pozycjami względem pozostałych. Ze względu na to, że LVF wykorzystuje do oceny istotności kryteriów współczynnik niespójności, daje ona kolejność cech w porządku rosnącym (najbardziej istotne cechy mają najniższe wartości niespójności).

[15] http://archive.ics.uci.edu/ml/index.html.

[16] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, *Modeling wine preferences by data mining from physicochemical properties*, „Decision Support Systems" 2009, nr 4, s. 547–553.

Tabela 1

Istotności cech uzyskane dla zbioru Car Evaluation

ReliefF	Cecha	6	4	1	2	5	3
	Istotność	0.3573	0.2908	0.2195	0.1944	0.0371	-0.0535
LVF	Cecha	1	6	4	2	5	3
	Niespójność	0.7	0.703	0.819	0.892	0.962	1
FCBF	Cecha	6	4	1	2	5	3
	Istotność	0.1879	0.1574	0.0602	0.046	0.0215	0.0028
CFS	Cecha	6	4	1	2	5	3
	Istotność	0.1879	0.1727	0.1352	0.1129	0.0946	0.0793
SA	Cecha	6	4	1	2	5	3
	Istotność	0.4334	0.3846	0.2455	0.2049	0.119	0.0567

Źródło: opracowanie własne.

W zbiorach Image Segmentation i Wine Quality widoczne są większe różnice w działaniu poszczególnych metod selekcji cech, jednak również w tych badaniach najbardziej od pozostałych odstają wyniki uzyskane za pomocą procedury LVF. W przypadku metody CFS łatwo zauważyć, że obliczone z istotności nie odzwierciedlają dokładnie kolejności w rankingu ważności cech. Rankingi dla tej procedury poprawiane są przez strategię *best first* i dopiero w ten sposób przyjmują kształt zbliżony do rankingów uzyskanych za pomocą pozostałych metod (z wyłączeniem LVF). Duże podobieństwo metod CFS i FCBF skutkuje tym, że cecha zajmująca najwyższą pozycję w rankingach utworzonych z wykorzystaniem tych dwóch metod zawsze ma taką samą lub bardzo zbliżoną wartość istotności. W rankingach utworzonych z wykorzystaniem metod ReliefF, FCBF i SA da się wyróżnić hipotetyczne podzbiory, w których wartości istotności cech są do siebie zbliżone w ramach podzbioru, zaś silnie odróżniają się od istotności cech w kolejnych podzbiorach. Dla zbioru danych Image Segmentation trzy podzbiory cech wyodrębnione w ten sposób z użyciem każdej z tych metod są do siebie podobne. Dla procedury SA podzbiór cech najistotniejszych zawierałby w tym przypadku atrybuty: 20, 12, 11, 18 i 14. Podzbiór taki dla procedury FCBF byłby dodatkowo rozszerzony o cechę 13, a w przypadku metody ReliefF zawierałby dodatkowo także atrybut 3. Z kolei podzbiór cech najmniej istotnych dla każ-

Tabela 2

Istotności cech uzyskane dla zbioru Image Segmentation

ReliefF	Cecha	20	13	18	11	12	14	3	17	16	15	19	2	9	7	5	6	10	8	4
	Istotność	0.2205	0.2201	0.2161	0.202	0.1961	0.1945	0.1904	0.1728	0.1618	0.1457	0.1451	0.0651	0.029	0.0231	0.0145	0.0077	0.0037	0.002	0
LVF	Cecha	12	3	20	7	2	10	16	9	4	5	6	8	11	13	14	15	17	18	19
	Niespójność	0.661	0.893	0.974	0.985	0.99	0.994	0.996	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
FCBF	Cecha	12	20	11	14	13	18	17	19	3	16	15	9	7	10	8	2	6	5	4
	Istotność	0.5629	0.5568	0.5212	0.5044	0.5026	0.501	0.4433	0.4322	0.4305	0.4186	0.3735	0.1828	0.175	0.1533	0.1375	0.0519	0.0153	0.0123	0
CFS	Cecha	12	20	3	14	17	19	15	13	9	16	11	2	7	18	6	5	10	8	4
	Istotność	0.563	0.677	0.705	0.709	0.71	0.711	0.709	0.705	0.7	0.695	0.692	0.689	0.684	0.68	0.675	0.67	0.664	0.656	0.545
SA	Cecha	20	12	11	18	14	19	13	17	3	16	15	7	9	8	10	6	2	5	4
	Istotność	0.9637	0.959	0.9407	0.9343	0.9317	0.9069	0.9019	0.8843	0.8828	0.8363	0.7828	0.5655	0.5644	0.5076	0.5043	0.2849	0.2519	0.1755	0

Źródło: opracowanie własne.

Tabela 3

Istotności cech uzyskane dla zbioru Wine Quality

ReliefF	Cecha	11	2	9	10	1	7	3	4	6	5	8
	Istotność	0.0166	0.0111	0.0103	0.0093	0.0084	0.0083	0.0082	0.0066	0.0064	0.0046	0.0041
LVF	Cecha	11	2	4	3	6	7	10	5	8	9	1
	Niespójn	0.493	0.539	0.569	0.617	0.679	0.743	0.805	0.854	0.888	0.91	0.915
FCBF	Cecha	11	8	5	7	3	6	2	4	9	1	10
	Istotność	0.09	0.0652	0.0488	0.0351	0.0347	0.0338	0.0324	0.0318	0.0117	0.0115	0.009
CFS	Cecha	11	8	5	3	2	6	7	4	1	9	10
	Istotność	0.09	0.0975	0.1026	0.106	0.1089	0.112	0.1118	0.1108	0.109	0.1073	0.1053
SA	Cecha	11	4	6	5	8	2	3	7	1	9	10
	Istotność	0.3545	0.3089	0.3043	0.2625	0.259	0.2484	0.2299	0.2234	0.1527	0.1334	0.0961

Źródło: opracowanie własne.

dej z metod zawierałby cechy: 2, 9, 7, 5, 6, 10, 8 i 4. Dla zbioru danych Wine Quality podzbiory wyodrębnione z użyciem każdej z trzech metod różniłyby się od siebie w większym stopniu. Analizując wyniki uzyskane metodami ReliefF, FCBF i SA, można zauważyć, że najbardziej odbiegają od pozostałych te, uzyskane z wykorzystaniem pierwszej procedury (ReliefF). Można więc przypuszczać, że najlepsze wyniki uzyskuje się z wykorzystaniem procedur FCBF i SA.

Podsumowanie

W uczeniu maszynowym redukcja wymiarowości przez selekcję cech stanowi ważne zagadnienie. Pozwala ona zmniejszyć komplikację zadania klasyfikacji, co niesie ze sobą wiele innych korzyści. W artykule przybliżono zagadnienia związane z selekcją cech. Zbadano również na trzech zbiorach danych działanie pięciu metod selekcji cech, należących do kategorii filtrów. W wyniku przeprowadzonych badań stwierdzono, że najbardziej wątpliwe wyniki określające istotność cech uzyskuje się z wykorzystaniem metody LVF. Można mieć również pewne zastrzeżenia odnośnie działania procedur CFS i ReliefF. Współczynniki istotności uzyskane w procedurze CFS nie odzwierciedlają dokładnie kolejności cech w rankingu utworzonym z wykorzystaniem tej metody. Z kolei procedura ReliefF daje rezultaty znacznie różniące się od wyników działania metod FCBF i SA, wobec czego przyjęto przypuszczenie, że spośród zbadanych metod najbardziej wiarygodne wyniki oferują procedury FCBF i SA. Należy jednak podkreślić, że zagadnienie to wymaga przeprowadzenia dalszych badań.

Literatura

Ahmad A., Dey L., *A feature selection technique for classificatory analysis*, „Pattern Recognition Letters" 2005, nr 26.

Chizi B., Maimon O., *Dimension Reduction and Feature Selection*, w: *Data Mining and Knowledge Discovery Handbook*, red. O. Maimon, L. Rokach, Springer, Nowy Jork 2010.

Cortez P., Cerdeira A., Almeida F., Matos T., Reis J., *Modeling wine preferences*

by data mining from physicochemical properties, „Decision Support Systems" 2009, nr 4.

Guyon I., *Practical Feature Selection: from Correlation to Causality*, w: *Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security*, red. F. Fogelman-Soulié, D. Perrotta, J. Piskorski, R. Steinberger, IOS Press, Amsterdam.

Hall M.A., *Correlation-based Feature Selection for Discrete and Nume*ric C*lass Machine Learning,* ICML '00 Proceedings of the 17th International Conference on Machine Learning 2000.

Hall M.A., Holmes G., *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*, „IEEE Transactions on Knowledge and Data Engineering" 2003, nr 3.

Hall M.A., Smith L.A., *Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper*, Proceedings of the 12th International Florida Artificial Intelligence Research Society Conference 1999, s. 235–239.

Hand D., Mannila H., Smyth D., *Eksploracja danych*, WNT, Warszawa 2005. http://archive.ics.uci.edu/ml/index.html.

Kannan S.S., Ramaraj N., *A novel hybrid feature selection via Symmetrical Uncerteinty ranking based local memetic search algorithm*, „Knowledge-Based Systems" 2010, nr 23.

Kira K., Rendell L.A., *A Practical Approach to Feature Selection*, ML92 Proceedings of the 9th international workshop on Machine learning 1992.

Kononenko I., Hong S.J., *Attribute Selection for Modelling*, „Future Generation Computer Systems" 1997, nr 2–3.

Liu H., Setiono R., *A Probabilistic Approach to Feature Selection – A Filter Solution*, Proceedings of the 13th International Conference on Machine Learning ICML'96.

Liu H., Yu L., Motoda H., *Feature Extraction, Selection, and Construction*, w: *The Handbook of Data Minin*g, red. N. Ye, Lawrence Erlbaum Associates, Mahwah 2003.

Michalak K., Kwaśnicka H.*, Correlation-based feature selection strategy in classification problems*, „International Journal of Applied Mathematics and Computer Science" 2006, nr 4.

Witten I.H., Frank E., *Data Mining. Practical Machine Learning Tools and Techniques*, Elsevier, San Francisco 2005.

Yu L., Liu H., *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, Proceedings of The 20th International Conference on Machine Leaning 2003.

THE REDUCTION OF ASSESSMENTS AND FEATURES SELECTION IN TASKS OF CLASSIFICATION AND THE REGRESSION WITH USING MACHINE LEARNING

Summary

Machine learning is being used in tasks of the regression and classification. In the field of classification a multidimensional of classified objects is one of essential problems. Classification is held on the basis of the value of features. These features are reflecting dimensions of the object subjected to the classification. In the article, applied algorithms were introduced selection of features which let reduce a problem "curses of dimensionality".

Keywords: machine learning, feature selection, "curses of dimensionality"

Translated by Paweł Ziemba