METODY INŻYNIERII WIEDZY

Metoda

K Najbliższych Sąsiadów

K-Nearest Neighbours

(KNN)

Adrian Horzyk

Akademia Górniczo-Hutnicza

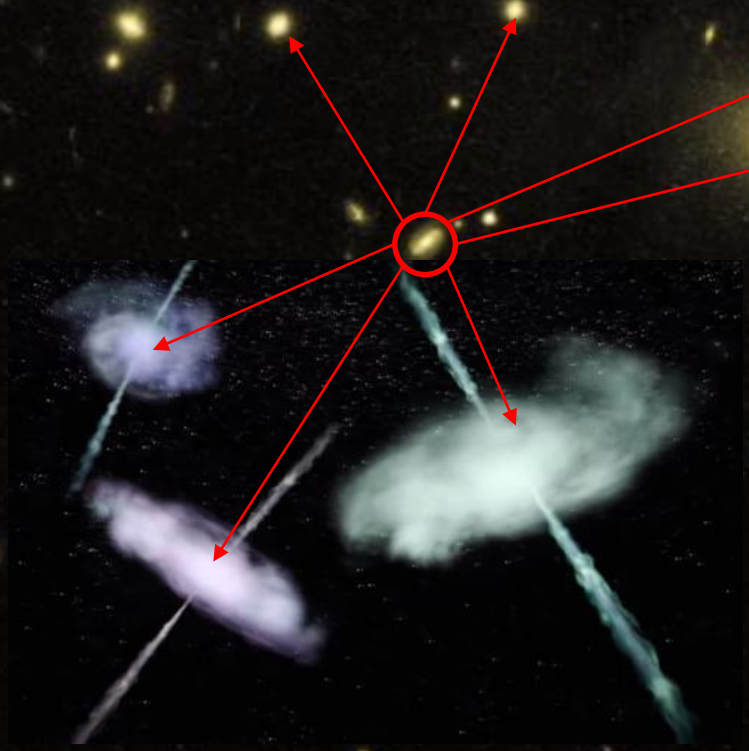
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Katedra Automatyki i Inżynierii Biomedycznej

Laboratorium Biocybernetyki

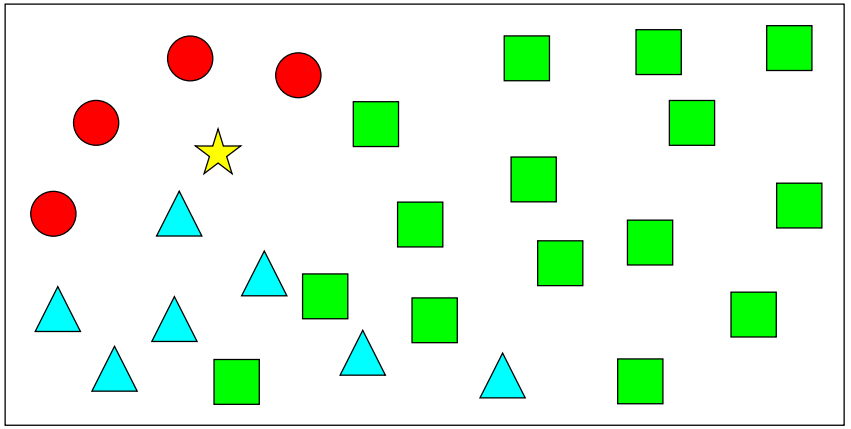
30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk



WSTĘP

- ✓ **Metoda K Najbliższych Sąsiadów (k-Nearest Neighbors) należy do grupy algorytmów leniwych (*lazy algorithms*), czyli takich, które nie tworzą wewnętrznej reprezentacji wiedzy o problemie na podstawie danych uczących, lecz szukają rozwiązania dopiero w momencie pojawienia się wzorca testowego do klasyfikacji. Metoda przechowuje wszystkie wzorce uczące, względem których wyznacza odległość wobec wzorca testowego.**

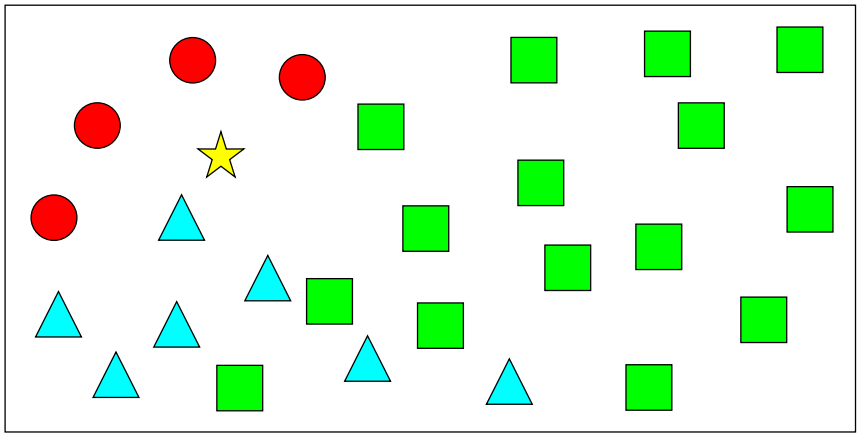


- ✓ **Do której klasy należy gwiazdka: kółeczek, trójkątów czy kwadratów?**

- ✓ **Istnieje też grupa metod gorliwych (*eager learning algorithms*).**
  Należą tutaj takie algorytmy uczące, które wewnętrzną reprezentację danych uczących (swoisty rodzaj wiedzy o problemie).
  Następnie w momencie pojawienia się wzorca testującego korzystając z tej wewnętrznej reprezentacji wiedzy dokonują jego klasyfikacji.
  Do tej grupy metod należą wszystkie rodzaje sieci neuronowych oraz systemy rozmyte, drzewa decyzyjne i wiele innych. Po zakończonej nauce (adaptacji) modelu, dane uczące mogą zostać usunięte.

ZBIÓR WZORCÓW UCZĄCYCH

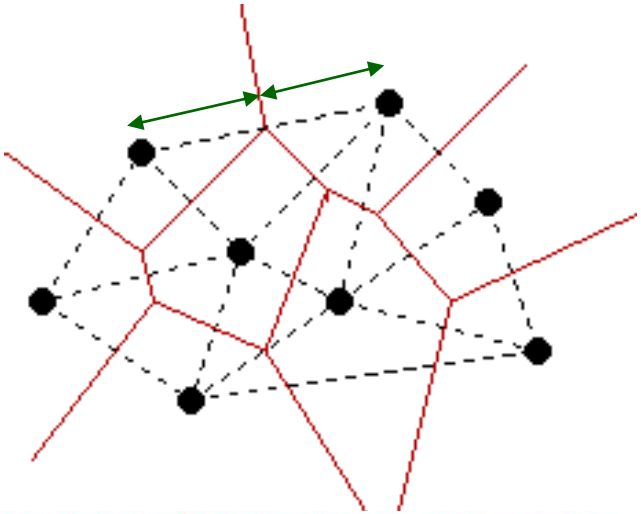
- ✓ **Zbiór wzorców uczących (*learning patterns, training examples*)** składa się ze zbioru par $<\mathbf{x}^i, y^i>$, gdzie $\mathbf{x}^i$ jest zbiorem parametrów $\mathbf{x}^i = \{x^i_1, \ldots, x^i_n\}$ definiujących obiekty (zwykle w postaci wektora lub macierzy danych), zaś y^i jest wartością przewidywaną/powiązaną/skojarzoną, np. indeksem lub nazwą klasy, do której obiekt $\mathbf{x}^i$ należy i którą razem z innymi obiektami tej klasy definiuje.

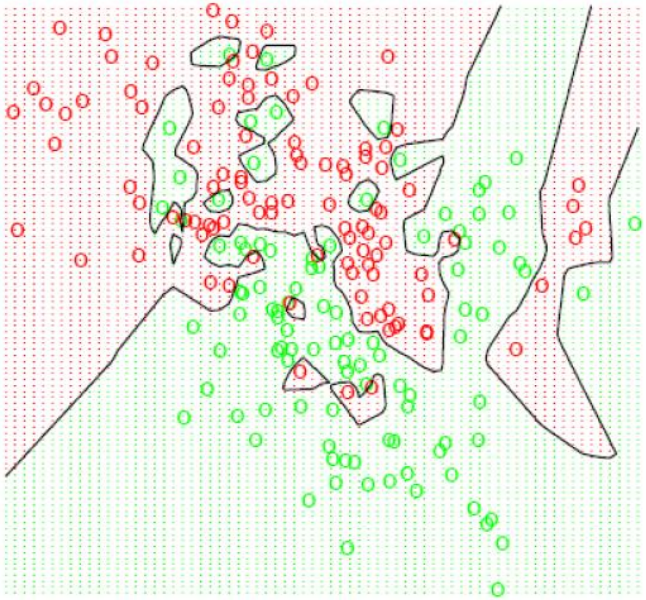


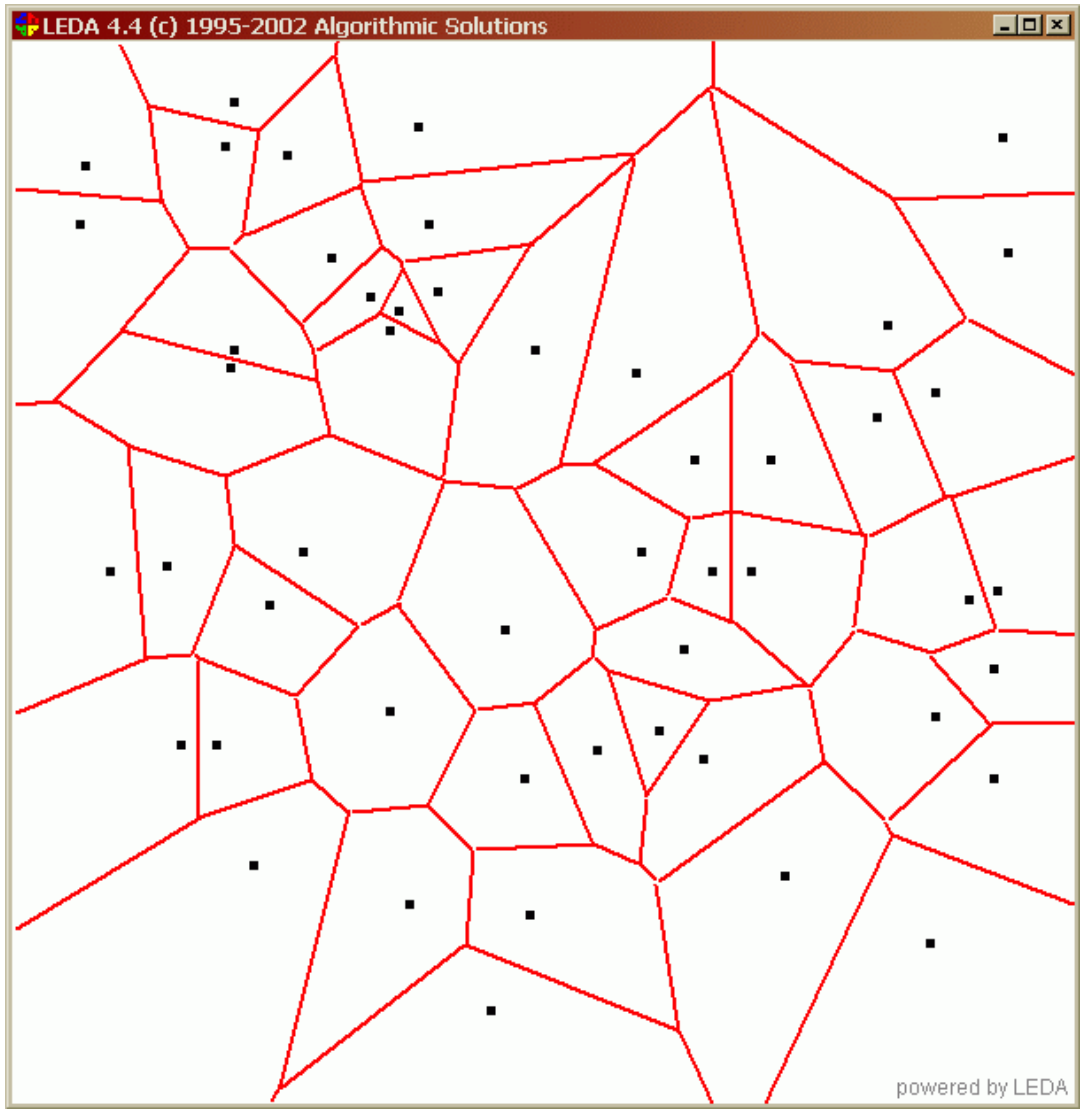
- ✓ Na rysunku mamy obiekty należące do 3 klas: kółeczka, trójkąciki i kwadraciki.
- ✓ Gwiazdka jest nowym obiektem, który chcemy sklasyfikować, czyli przyporządkować go do jednej z istniejących klas obiektów.
- ✓ Do którego zbioru należy gwiazdka? Co nam podpowiada intuicja?
- ✓ Można np. badać odległość gwiazdki od pozostałych obiektów, dla których klasa jest znana, korzystając z jednej ze znanych metryk, np. odległości Euklidesa: $\|x - x^k\|_2 = \sqrt{\sum_{j=0}^{J} (x_j - x_j^k)^2}$

DIAGRAMY VORONOI

- ✓ **Diagramy Voronoi (*Voronoi diagrams*) ilustrują obszary przyciągania (*attraction areas*) do najbliższych pojedynczych elementów w przestrzeni:**

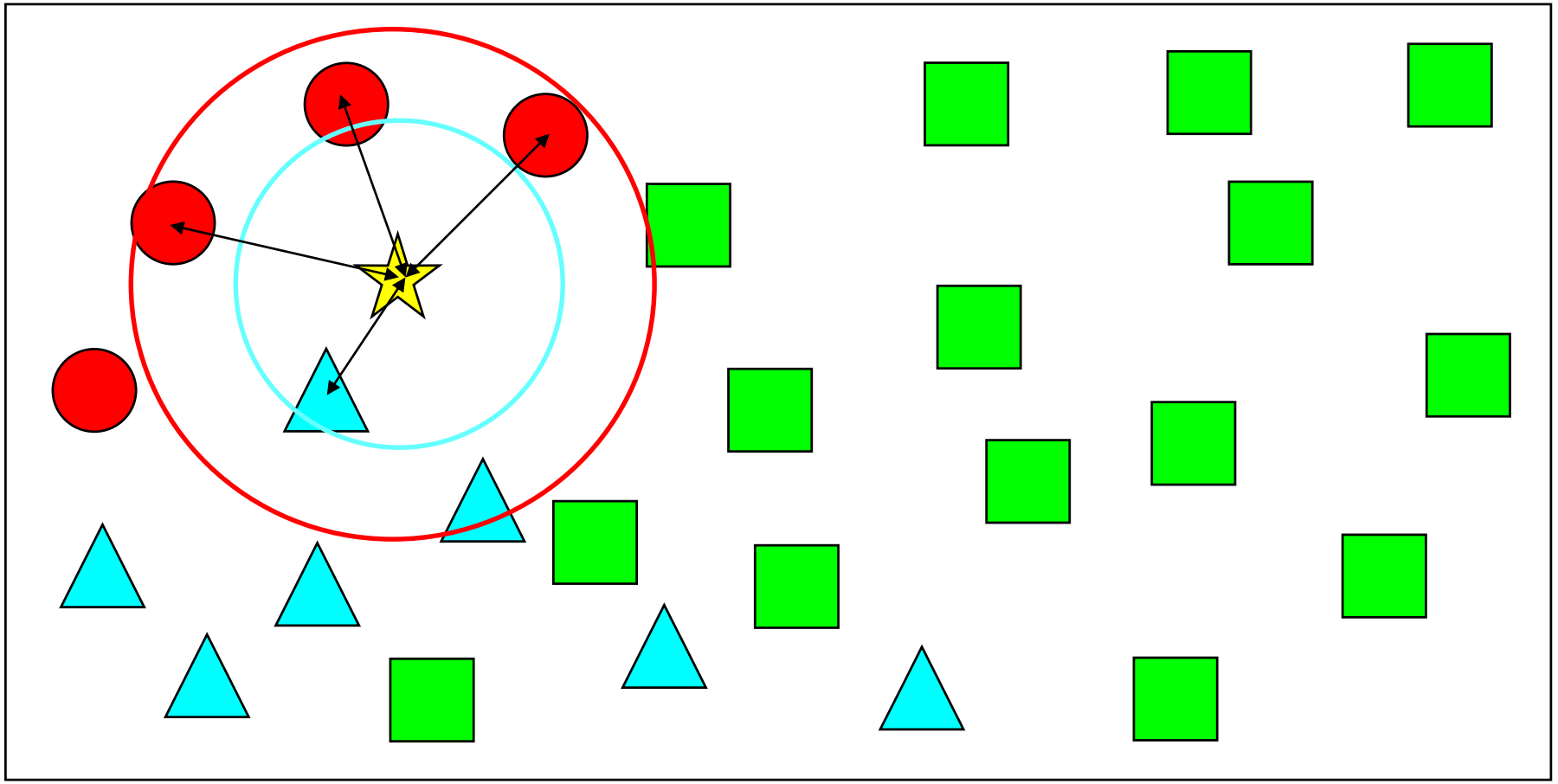






METODA K NAJBLIŻSZYCH SĄSIADÓW

- ✓ **Metoda k Najbliższych Sąsiadów (k-Nearest Neighbors) wyznacza k sąsiadów, do których badany element (gwiazdka) ma najbliżej dla wybranej metryki (np. Euklidesowej), a następnie wyznacza wynik w oparciu o głos większości:**



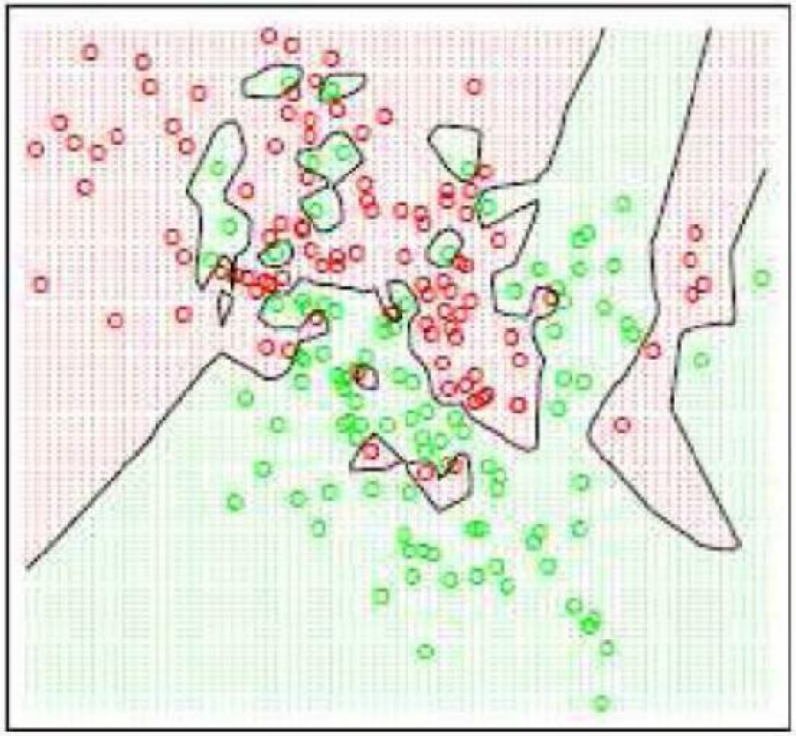
Do którego zbioru należy gwiazdka?

Wynik działania zależy to od tego, ilu k najbliższych sąsiadów weźmiemy pod uwagę.

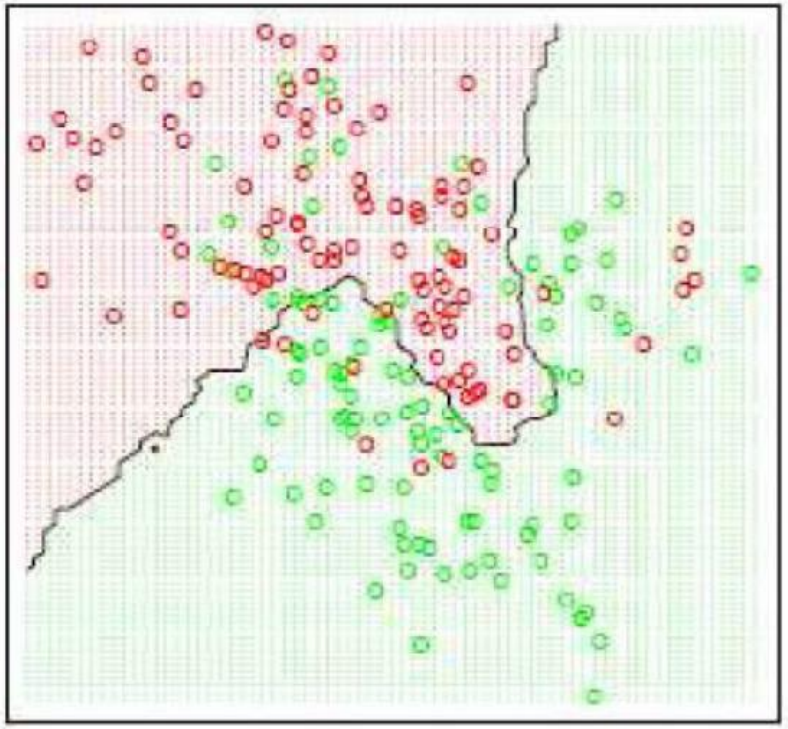
METODA K NAJBLIŻSZYCH SĄSIADÓW

- ✓ **Metoda k Najbliższych Sąsiadów (k-Nearest Neighbors)** daje różne wyniki w postaci obszarów przyciągania, co determinuje wynik klasyfikacji:

K=1



K=15



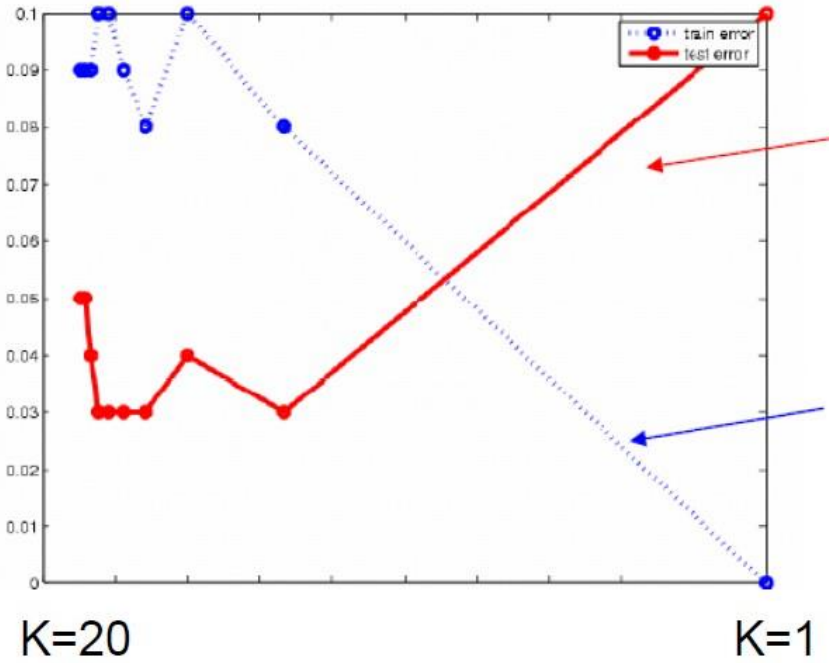
Figures from Hastie, Tibshirani and Friedman (Elements of Statistical Learning)

- ✓ Większe wartości k umożliwiają wygładzenie obszarów podziału, usunięcie szumu i artefaktów, lecz również prowadzą do błędów w klasyfikacji rzadszych wzorców. Pojawia się problem ich poprawnej dyskryminacji i uogólniania.

Dobieranie wartości k

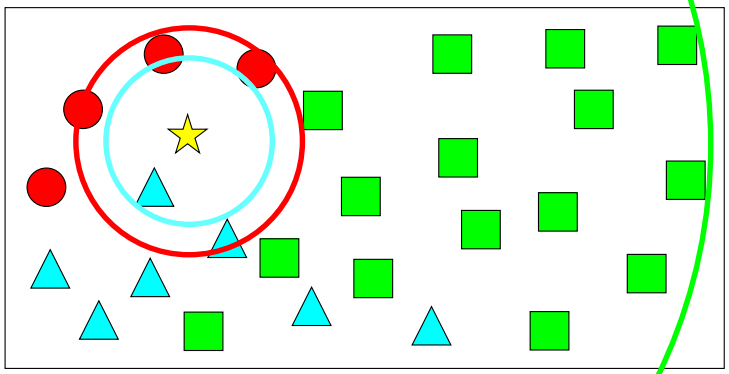
- ✓ Jeślibyśmy wzięli pod uwagę k=N, gdzie N to ilość wszystkich elementów zbioru wzorców uczących, wtedy zawsze wynik klasyfikacji będzie określony przez najliczniej reprezentowaną klasę w tym zbiorze uczących, a więc w tym przykładzie: kwadraciki.

Przykładowy wykres zależności wyników adaptacji od wartości k:



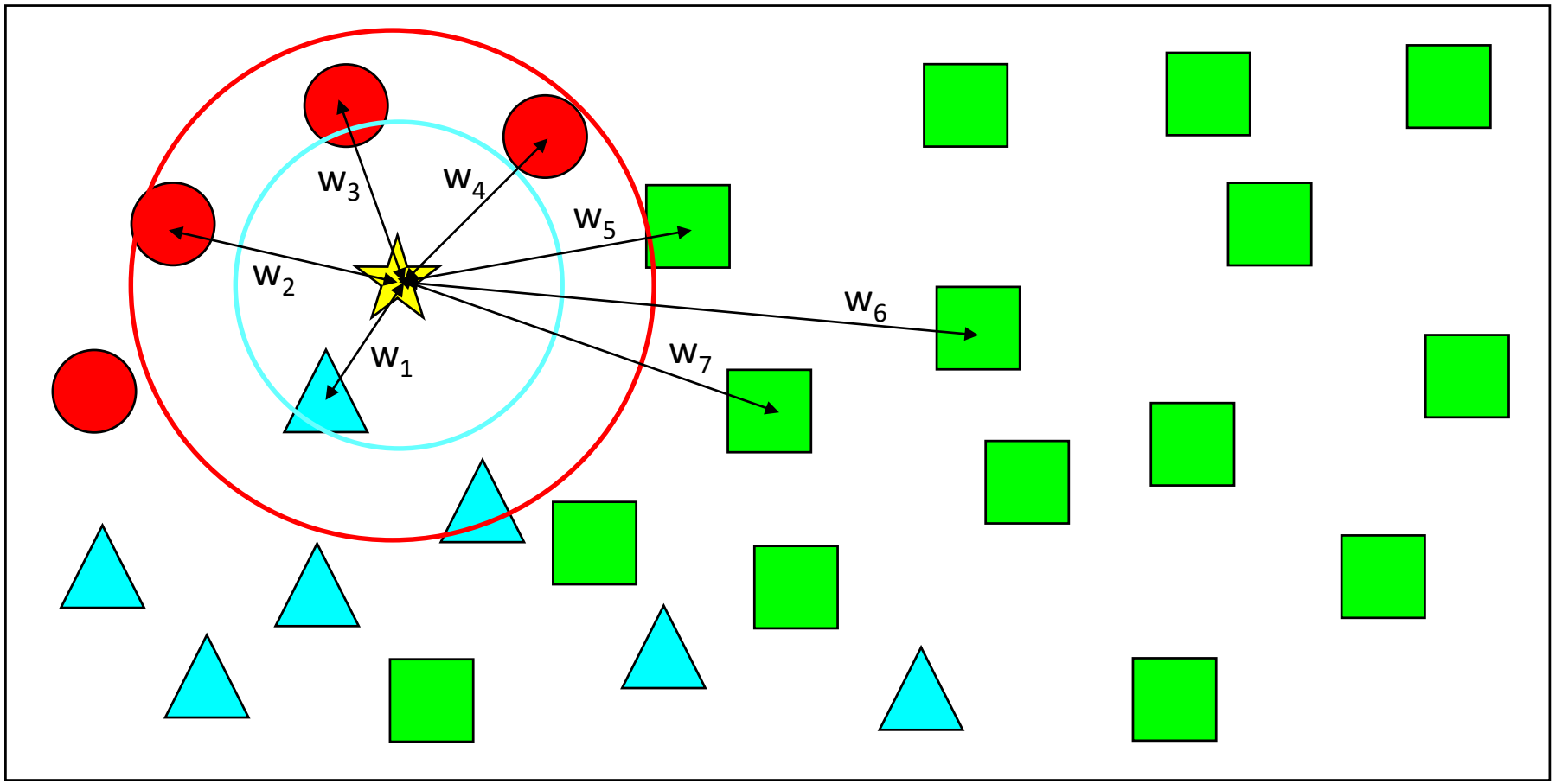
BŁĄD TESTU

BŁĄD UCZENIA



MODYFIKACJE METODY K NAJBLIŻSZYCH SĄSIADÓW

Metoda Ważonych Odległości Najbliższych Sąsiadów (Distance Weighted Nearest Neighbors) prowadzi do głosowania na temat klasyfikacji gwiazdki biorąc pod uwagę k najbliższych sąsiadów, lecz ich **głosy są ważone w zależności od ich odległości** (dla wybranej metryki) do gwiazdki: im dalej jest głosujący wzorzec tym ma mniejszą wagę. A więc wzorce położone najbliżej będą miały największy wpływ na wynik klasyfikacji. Można wziąć też pod uwagę wszystkie wzorce i ich odległości od wzorca klasyfikowanego, lecz wtedy trzeba wagę podzielić przed liczność klasy, do której należy, gdyż w odwrotnym przypadku uprzywilejowane byłyby klasy najbardziej liczne!



WAŻENIE ODLEGŁOŚCI ORAZ NORMALIZACJA WZGLĘDEM LICZNOŚCI POSZCZEGÓLNYCH KLAS

Jak dobrać wagi dla Metody Ważonych Odległości Najbliższych Sąsiadów?

Należy wziąć pod uwagę taką zależność wagi od odległości, żeby waga malała wraz ze zwiększającą się odległością. Możliwości jest kilka (proszę poeksperymentować), np.:

$$w = \frac{d_{max} - d}{d_{max} \cdot N_{class}}$$

$$w = \frac{1}{d \cdot N_{class}}$$

$$w = \frac{1}{\sqrt[m]{d} \cdot N_{class}}$$

Gdzie:

d – odległość Euklidesa badanego wzorca od wzorca uczącego

m – stopień pierwiastka: 2, 3, 4…

d_{max} – to maksymalna odległość wzorca spośród k-najbliższych (badanych)

N_{class} – to liczność klasy „class”, do której należy wzorzec uczący względem którego wyznaczamy odległość i który bierze udział w tym ważonym odległością „głosowaniu”.

KONKURS NA NAJLEPSZĄ ZMODYFIKOWANĄ METODĘ KNN!

TRUDNOŚCI METOD K NAJBLIŻSZYCH SĄSIADÓW

- ✓ **Metody Najbliższych Sąsiadów ponoszą karę za swoje „lenistwo”, związane z brakiem budowy modelu generalizującego wzorce uczące. Wymagają przeglądania w pętlach wszystkich wzorców uczących dla wszystkich wymiarów (złożoność wielomianowa).**
- ✓ **Nawet w przypadku poindeksowania wzorców względem każdego wymiaru/parametru w bazie danych, trzeba przejść po każdym z wymiarów, w celu określenia najbliższych sąsiadów, co też jest czasochłonne w zależności od wyboru k. W metodach sprawdzających ważone odległości do wszystkich wzorców i tak trzeba przejść po nich wszystkich każdorazowo.**
- ✓ **W przypadku dużej ilości wzorców lub dużego wymiaru danych klasyfikacja realizacja klasyfikacji wiąże się z ogromnym narzutem czasowym. Od tego wolne są inne modele, które dla wzorców uczących najpierw budują model (zwykle jednokrotnie pewnym nakładem czasowym), lecz potem klasyfikacja przebiega już bardzo szybko!**
- ✓ **Metoda jest ponadto wrażliwa na dane zaszumione lub błędne, np. w odniesieniu do niektórych cech wzorców (*noisy features*). Mogą one zmienić wynik klasyfikacji.**
- ✓ **Metoda jest wrażliwa na różnoliczność wzorców reprezentujących poszczególne klasy.**
- ✓ **Metoda k-NN natomiast stosunkowo dobrze działa dla dużej ilości klas.**

PRÓBY ROZWIĄZANIA:

- ✓ **Adaptacja sposobu określania odległości poprzez adaptacyjne ważenie cech, odległości, wartości k.**

JAK ZBUDOWAĆ DOBRZE DZIAŁAJĄCY KLASYFIKATOR KNN?

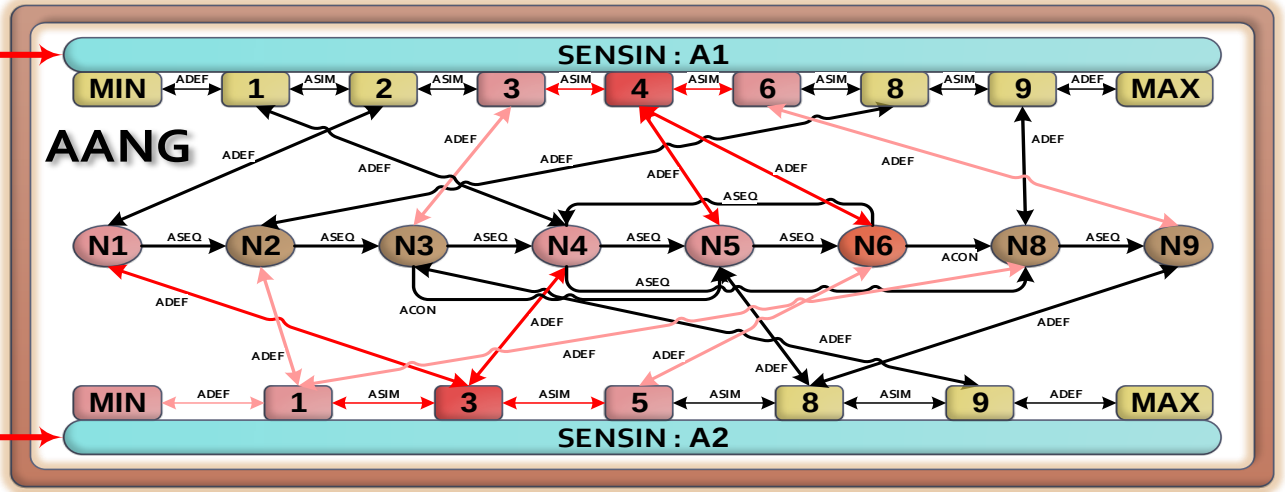
- ✓ **Sprawdzić, czy klasy definiowane są przez podobną ilość wzorców?**
  Jeśli nie są, niezbędna będzie normalizacja ilości sąsiadów przez liczność poszczególnych klas w celu uzyskania wiarygodnych wyników.
- ✓ **Czy zagęszczenie danych w przestrzeni jest równomierne?**
  Jeśli nie jest, warto wykorzystać głosowanie z wykorzystaniem ważenia tych głosów poprzez odległości głosującego wzorca do wzorca klasyfikowanego, w celu zwiększenia wiarygodności.
- ✓ **Jak dobrać odpowiednie k?**
  Najlepiej przeprowadzić dla znanych danych uczących metodą walidacji krzyżowej sprawdzenie, które k dla danego zbioru danych uczących będzie najlepsze.
  Jeśli dane reprezentujące poszczególne klasy nie są mniej więcej równoliczne, wtedy dobieranie dużych k jest ryzykowne, gdyż preferowane są klasy liczniejsze.
  Jeśli dane uczące są zaszumione, częściowo błędne lub odstające, wtedy stosowanie k=1 odradza się, gdyż jest duże niebezpieczeństwo błędnej klasyfikacji.
- ✓ **Najbezpieczniej zastosować podejście ważone z normalizacją względem liczności klas oraz dokonać adaptacyjnego doboru wartości k na postawie walidacji krzyżowej dla określonego zbioru uczącego, testując różne wartości od k=1 do k kilkanaście. Większe wartości k niż kilkanaście rzadko dają lepsze rezultaty.**

MODYFIKACJE METODY K NAJBLIŻSZYCH SĄSIADÓW

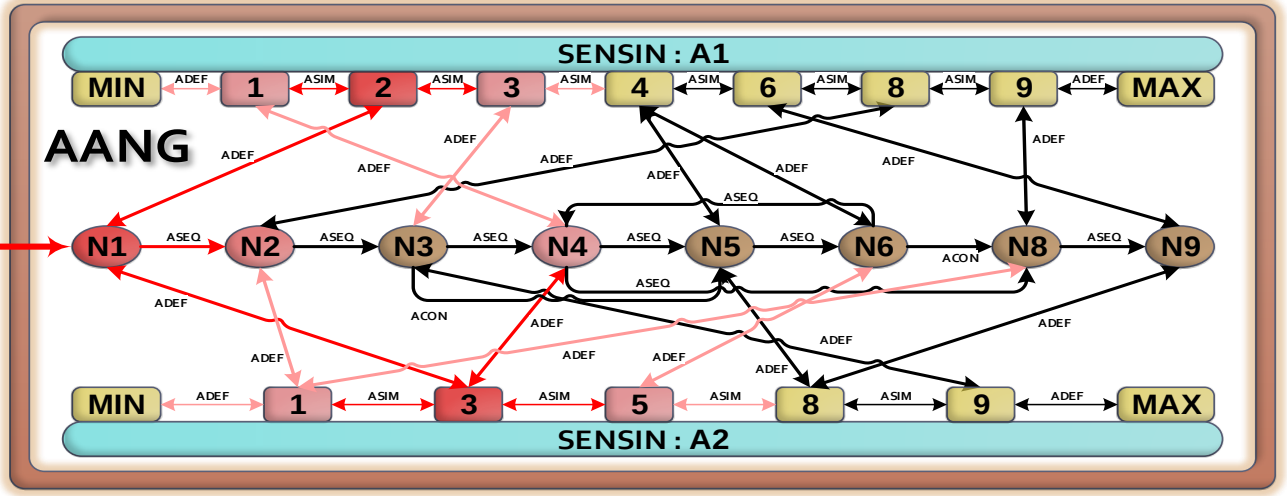
- ✓ **Korzystając np. z asocjacyjnego sortowania ASSORT oraz neuronowych struktur AANG można by było bardzo szybko wyznaczyć najbliższych sąsiadów na podstawie podobieństwa do prezentowanego wzorca na wejściu, lecz po co, skoro sama sieć AANG również umożliwia klasyfikację, zwykle nie gorszą niż k-NN, a w dodatku tworzy model na podstawie danych uczących, co znacznie przyspiesza końcowy proces klasyfikacji.**

ZEWNĘTRZNE POBUDZANIE WEJŚĆ SENSORYCZNYCH

4 3



ZEWNĘTRZNE POBUDZANIE NEURONU



PORÓWNANIE WYNIKÓW KLASYFIKACJI AANG i kNN

CONGRESS VOTING (435 wzorców, 16 parametrów	UCZEN	WALID	IONOSPHERE (29289 wzorców, 34 parametrów)	UCZEN	WALID
Algorytm i parametry	Q_COR (avg)	Q_COR (avg)	Algorytm (parametry)	Q_COR (avg)	Q_COR (avg)
NAJLEPSZY WYNIK:	**100,00**	**96,55**	**NAJLEPSZY WYNIK:**	**100,00**	**94,30**
SVM Multiclass Linear (cost=50;eps=0,1)	96,78	96,55	SVM Multiclass RBF (cost=10;sigma=1;eps=0,1)	100,00	94,30
SVM Multiclass RBF (cost=10;sigma=5;eps=0,1)	98,60	96,08	SVM Multiclass RBF (cost=10;sigma=5;eps=0,1)	97,06	94,30
SVM Multiclass Linear (cost=1;eps=0,1)	97,16	96,08	**ASONN (stopień dyskryminacji=2;rozmycie brzegów=2)**	99,27	91,75
SVM Multiclass Linear (cost=20;eps=0,001)	97,47	96,08	**AANG ()**	100,00	91,46
SVM Multiclass Linear (cost=100;eps=0,1)	96,65	96,07	**ASONN (stopień dyskryminacji=3;rozmycie brzegów=1)**	99,21	90,87
RBFN (k=15;funk-radial=1;met.wybor.centr=2)	95,96	96,07	**ASONN (stopień dyskryminacji=2;rozmycie brzegów=1)**	99,11	90,87
SVM Multiclass Linear (cost=20;eps=0,1)	97,16	95,86	RBFN (k=15;funk-radial=1;met.wybor.centr=2)	90,88	90,32
SVM Multiclass Linear (cost=10;eps=0,1)	97,04	95,85	RBFN(k=15;funk-radial=2;met.wybor.centr=2)	91,04	90,31
RBFN (k=20;funk-radial=1;met.wybor.centr=2)	96,53	95,85	SVM Multiclass RBF (cost=10;sigma=10;eps=0,1)	94,08	90,30
SVM Multiclass Linear (cost=20;eps=0,01)	97,47	95,85	SVM MulticlassP (cost=10;d=4;eps=0,01)	96,96	90,06
SVM Multiclass RBF (cost=10;sigma=10;eps=0,1)	96,73	95,39	SVM Multiclass Linear (cost=10;eps=0,1)	93,95	90,02
ASONN (stopień dyskryminacji=1;rozmycie brzegów=2)	100,00	95,18	**ASONN (stopień dyskryminacji=1;rozmycie brzegów=1)**	98,54	89,75
RBFN(k=15;funk-radial=2;met.wybor.centr=2)	95,94	95,17	RBFN (k=20;funk-radial=2;met.wybor.centr=1)	91,52	89,74
RBFN (k=10;funk-radial=1;met.wybor.centr=2)	95,12	94,92	RBFN (k=15;funk-radial=2;met.wybor.centr=1)	91,04	89,73
ASONN (stopień dyskryminacji=2;rozmycie brzegów=1)	100,00	94,72	SVM Multiclass Linear (cost=50;eps=0,1)	92,72	89,73
ASONN (stopień dyskryminacji=2;rozmycie brzegów=2)	100,00	94,72	**ASONN (stopień dyskryminacji=3;rozmycie brzegów=3)**	99,15	89,17
SONN-3 (Adrian Horzyk)()	100,00	94,71	RBFN(k=20;funk-radial=2;met.wybor.centr=2)	91,83	89,17
ASONN (stopień dyskryminacji=1;rozmycie brzegów=1)	100,00	94,49	**ASONN (stopień dyskryminacji=3;rozmycie brzegów=2)**	99,21	89,17
ASONN (stopień dyskryminacji=3;rozmycie brzegów=3)	100,00	94,48	SVM Multiclass Linear (cost=20;eps=0,1)	93,80	88,88
ASONN (stopień dyskryminacji=Auto;rozmycie brzegów=1)	100,00	94,48	SVM Multiclass Linear (cost=20;eps=0,001)	94,87	88,88
ASONN (stopień dyskryminacji=3;rozmycie brzegów=2)	100,00	94,48	**ASONN (stopień dyskryminacji=1;rozmycie brzegów=2)**	98,54	88,61
ASONN (stopień dyskryminacji=3;rozmycie brzegów=1)	100,00	94,48	RBFN (k=10;funk-radial=1;met.wybor.centr=2)	89,97	88,32
RBFN (k=10;funk-radial=2;met.wybor.centr=2)	94,23	94,01	SVM Multiclass RBF (cost=10;sigma=0,5;eps=0,1)	100,00	88,31
RBFN (k=5;funk-radial=1;met.wybor.centr=2)	94,05	93,32	SVM Multiclass Linear (cost=20;eps=0,01)	94,75	88,02
PNN Simple (sigma=0,1)	100,00	93,10	RBFN (k=20;funk-radial=1;met.wybor.centr=2)	92,15	87,75
K nn (k=1)	100,00	92,88	SVM Multiclass Linear (cost=1;eps=0,1)	93,16	87,74
RBFN (k=10;funk-radial=2;met.wybor.centr=1)	92,95	92,86	SVM Multiclass Linear (cost=100;eps=0,1)	92,09	87,18
AANG ()	100,00	92,63	K nn (k=1)	100,00	86,90
K nn (k=3)	92,49	92,39	RBFN (k=10;funk-radial=2;met.wybor.centr=1)	88,54	86,89
PNN Simple (sigma=1)	98,75	92,18	**SONN-3 (Adrian Horzyk)()**	97,28	86,62
K nn (k=5)	91,16	91,93	**ASONN (stopień dyskryminacji=Auto;rozmycie brzegów=1)**	98,96	86,59
K nn (k=10)	91,24	90,32	PNN Simple (sigma=1)	94,30	86,33
RBFN (k=5;funk-radial=2;met.wybor.centr=2)	90,96	89,87	SVM MulticlassP (cost=10;d=3;eps=0,01)	97,88	86,02
K nn (k=15)	89,22	89,20	K nn (k=3)	85,82	83,48
PNN Simple (sigma=100)	89,35	88,95	RBFN (k=5;funk-radial=1;met.wybor.centr=2)	84,33	83,25
PNN Simple (sigma=10)	89,40	88,95	RBFN (k=10;funk-radial=2;met.wybor.centr=2)	86,99	82,90
K nn (k=30)	88,48	88,94	K nn (k=5)	82,40	80,35
SVM Multiclass RBF (cost=10;sigma=1;eps=0,1)	100,00	85,98	SVM MulticlassP (cost=10;d=6;eps=0,01)	81,16	80,35
SVM Multiclass RBF (cost=10;sigma=0,5;eps=0,1)	100,00	78,39	RBFN (k=5;funk-radial=2;met.wybor.centr=1)	80,41	80,33

PORÓWNANIE WYNIKÓW KLASYFIKACJI AANG i kNN

WINE (178 wzorców, 13 parametrów)	UCZEN	WALID
Algorytm i parametry	Q_COR (avg)	Q_COR (avg)
NAJLEPSZY WYNIK:	**100,00**	**99,44**
ASONN (stopień dyskryminacji=2; rozmycie brzegów=2)	100,00	99,44
RBFN(k=15; funkcja radialna=1; metoda wyboru centrów=2)	99,69	99,41
ASONN (stopień dyskryminacji=2; rozmycie brzegów=1)	100,00	98,89
ASONN (stopień dyskryminacji=3; rozmycie brzegów=1)	100,00	98,89
ASONN (stopień dyskryminacji=3; rozmycie brzegów=2)	100,00	98,89
ASONN (stopień dyskryminacji=3; rozmycie brzegów=3)	100,00	98,89
RBFN(k=20; funkcja radialna=1; metoda wyboru centrów=2)	99,94	98,30
ASONN (stopień dyskryminacji=0; rozmycie brzegów=1)	100,00	97,78
AANG ()	100,00	97,19
ASONN (stopień dyskryminacji=1; rozmycie brzegów=1)	100,00	96,67
RBFN(k=10; funkcja radialna=1; metoda wyboru centrów=2)	98,50	96,63
ASONN (stopień dyskryminacji=1; rozmycie brzegów=2)	100,00	96,11
SVM Multiclass Linear(cost=20; eps=0,001)	97,63	95,56
SONN-3 ()	100,00	91,08
SVM Multiclass Linear(cost=20; eps=0,01)	93,95	90,49
RBFN(k=5; funkcja radialna=1; metoda wyboru centrów=2)	82,52	84,38
SVM Multiclass RBF(cost=10; sigma=10; eps=0,1)	100,00	78,04
K nn(k=1)	100,00	77,52
PNN Simple(sigma=1)	100,00	75,82
PNN Simple(sigma=10)	90,82	75,78
PNN Simple(sigma=100)	73,91	72,97
RBFN(k=10; funkcja radialna=2; metoda wyboru centrów=2)	69,66	69,61
RBFN(k=5; funkcja radialna=2; metoda wyboru centrów=2)	69,85	68,46
RBFN(k=15; funkcja radialna=2; metoda wyboru centrów=2)	67,80	67,35
RBFN(k=15; funkcja radialna=2; metoda wyboru centrów=1)	67,79	66,83
K nn(k=3)	67,48	66,27
K nn(k=5)	66,29	66,24
K nn(k=30)	62,66	66,18
SVM Multiclass RBF(cost=10; sigma=5; eps=0,1)	100,00	65,75
RBFN(k=5; funkcja radialna=2; metoda wyboru centrów=1)	69,66	65,69
K nn(k=15)	64,80	65,13
RBFN(k=20; funkcja radialna=2; metoda wyboru centrów=1)	69,29	65,10
K nn(k=10)	68,85	64,09
SVM Multiclass RBF(cost=10; sigma=0,5; eps=0,1)	100,00	39,93
SVM Multiclass RBF(cost=10; sigma=1; eps=0,1)	100,00	39,93
SVM Multiclass Linear(cost=1; eps=0,1)	33,15	33,14
SVM Multiclass Linear(cost=10; eps=0,1)	33,15	33,14
SVM Multiclass Linear(cost=100; eps=0,1)	33,15	33,14

IRIS (150 wzorców, 4 parametry)	UCZEN	WALID
Algorytm i parametry	Q_COR (avg)	Q_COR (avg)
NAJLEPSZY WYNIK:	**100,00**	**98,00**
RBFN(k=15;funkcja radialna=1;metoda wyboru centrów=2)	98,89	98,00
RBFN(k=20;funkcja radialna=1;metoda wyboru centrów=2)	98,44	97,33
SVM Multiclass RBF(cost=10;sigma=5;eps=0,1)	97,41	96,67
K nn(k=1)	100,00	96,00
K nn(k=10)	93,93	96,00
K nn(k=15)	92,07	96,00
RBFN(k=10;funkcja radialna=2;metoda wyboru centrów=2)	97,04	96,00
PNN Simple(sigma=0,1)	100,00	96,00
ASONN (stopień dyskryminacji=1;rozmycie brzegów=1)	100,00	95,33
ASONN (stopień dyskryminacji=1;rozmycie brzegów=2)	100,00	95,33
SVM Multiclass RBF(cost=10;sigma=0,5;eps=0,1)	99,26	95,33
SVM Multiclass RBF(cost=10;sigma=1;eps=0,1)	98,44	95,33
SVM Multiclass RBF(cost=10;sigma=10;eps=0,1)	95,63	95,33
RBFN(k=5;funkcja radialna=2;metoda wyboru centrów=2)	96,07	95,33
K nn(k=5)	95,63	94,67
SVM Multiclass Linear(cost=10;eps=0,1)	94,81	94,67
SVM Multiclass Linear(cost=20;eps=0,1)	95,56	94,67
SVM Multiclass Linear(cost=20;eps=0,001)	95,85	94,67
RBFN(k=5;funkcja radialna=2;metoda wyboru centrów=1)	95,63	94,67
RBFN(k=10;funkcja radialna=1;metoda wyboru centrów=2)	97,78	94,67
AANG ()	100,00	94,00
ASONN (stopień dyskryminacji=0;rozmycie brzegów=1)	100,00	94,00
ASONN (stopień dyskryminacji=2;rozmycie brzegów=1)	100,00	94,00
ASONN (stopień dyskryminacji=2;rozmycie brzegów=2)	100,00	94,00
ASONN (stopień dyskryminacji=3;rozmycie brzegów=1)	100,00	94,00
ASONN (stopień dyskryminacji=3;rozmycie brzegów=2)	100,00	94,00
ASONN (stopień dyskryminacji=3;rozmycie brzegów=3)	100,00	94,00
K nn(k=3)	94,52	94,00
SVM Multiclass Linear(cost=100;eps=0,1)	94,37	94,00
PNN Simple(sigma=1)	94,59	94,00
SVM Multiclass Linear(cost=1;eps=0,1)	93,85	93,33
SVM Multiclass Linear(cost=20;eps=0,01)	95,41	93,33
RBFN(k=5;funkcja radialna=1;metoda wyboru centrów=2)	95,33	93,33
RBFN(k=15;funkcja radialna=2;metoda wyboru centrów=2)	96,37	92,67
SVM Multiclass Linear(cost=50;eps=0,1)	94,59	92,00
SONN-3 ()	99,85	91,33
PNN Simple(sigma=10)	91,48	90,67
PNN Simple(sigma=100)	91,48	90,67