PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Dorota Rozmus
Akademia Ekonomiczna w Katowicach

ZASTOSOWANIE MIAR POZYCYJNYCH DO BADANIA RELACJI MIĘDZY ZRÓŻNICOWANIEM A DOKŁADNOŚCIĄ KLASYFIKACJI W PODEJŚCIU ZAGREGOWANYM W TAKSONOMII

Streszczenie: Dotychczas podejście wielomodelowe z dużym powodzeniem stosowane było w dyskryminacji i regresji w celu podniesienia dokładności predykcji. W ostatnich latach analogiczne propozycje pojawiły się w taksonomii, aby zapewnić większą poprawność i stabilność wyników klasyfikacji.

Ważnym czynnikiem przyczyniającym się do sukcesu podejścia wielomodelowego jest zróżnicowanie elementów wchodzących w skład klasyfikacji zagregowanej. Zasadniczym celem badania jest próba zastosowania miar pozycyjnych w zaproponowanych dotąd miernikach zróżnicowania [Hadjitodorov, Kuncheva, Todorova 2006] do zbadania relacji, jakie zachodzą między poziomem zróżnicowania klasyfikacji składowych a jakością klasyfikacji zagregowanej w podejściu wielomodelowym w taksonomii oraz porównanie wyników z dotychczas stosowanymi sposobami badania tych relacji.

1. Wstęp

Dotychczas podejście wielomodelowe z dużym powodzeniem stosowane było w dyskryminacji i regresji w celu podniesienia dokładności predykcji. Idea tego podejścia polega na tym, że w pierwszym kroku budowane są liczne pojedyncze i różniące się między sobą modele, które następnie za pomocą różnych operatorów łączone są w model zagregowany. W klasyfikacji najczęściej stosowane jest głosowanie majoryzacyjne, a więc obiekt klasyfikowany jest do tej klasy, która najczęściej wskazywana była przez pojedyncze modele; natomiast w regresji najczęściej stosuje się uśrednianie wartości teoretycznych zmiennej y. Do najbardziej znanych metod agregacji należą: *bagging* [Breiman 1996] oparty na losowaniu prób boostrapowych oraz *boosting* [Freund 1990] polegający na nadawaniu wyższych wartości wag błędnie sklasyfikowanym obiektom.

W ostatnich latach analogiczne propozycje pojawiły się w taksonomii, aby zapewnić większą dokładność[1] i stabilność wyników klasyfikacji [Fred 2002; Fred,

[1] Przez dokładność klasyfikacji należy rozumieć zgodność uzyskanych rezultatów z rzeczywistą strukturą klas.

Jain 2002; Strehl, Ghosh 2002]. Zasadnicza idea tego podejścia polega na połączeniu wyników wielokrotnie przeprowadzonego grupowania. Zagadnienie łączenia w taksonomii można sformułować następująco: mając dane wyniki wielokrotnie przeprowadzonej klasyfikacji, określ zagregowany podział ostateczny. Liczne badania w tej dziedzinie ugruntowały już nowy obszar w tradycyjnej taksonomii. Istnieje kilka możliwości zastosowania idei podejścia zagregowanego w dziedzinie taksonomii:

1) łączenie wyników grupowania uzyskanych za pomocą różnych metod,
2) uzyskanie różniących się między sobą klasyfikacji z zastosowaniem różnych podzbiorów danych, np. przez losowanie bootstrapowe,
3) stosowanie różnych podzbiorów zmiennych,
4) zastosowanie wielokrotnie tego samego algorytmu z różnymi wartościami parametrów lub punktami startowymi (np. losowo wybranymi zalążkami skupień w metodzie *k*-średnich).

2. Miary zróżnicowania

Ważnym czynnikiem przyczyniającym się do sukcesu podejścia zagregowanego jest zróżnicowanie elementów wchodzących w skład klasyfikacji zagregowanej [Tsymbal, Pechenizkiy, Cunningham 2003; Kuncheva, Hadjitodorov 2004]. Agregacja identycznych składowych nie pozwoli na poprawę dokładności klasyfikacji. Jakkolwiek znalezienie odpowiedniej miary zróżnicowania w podejściu wielomodelowym w dyskryminacji okazało się zadaniem bardzo trudnym [Kuncheva 2003]. W ostatnich latach rozpatruje się także zagadnienie zróżnicowania w taksonomii. Fern i Brodley [2003] zauważają, że bardziej zróżnicowane klasyfikacje składowe dają większą poprawę dokładności klasyfikacji w porównaniu z mniej zróżnicowanymi. W literaturze zaproponowano wiele licznych miar zróżnicowania. W niniejszym badaniu wykorzystano ideę wskaźników zaproponowanych przez Hadjitodorova, Kunchevę i Todorovą [2006], którzy zaproponowali pięć mierników zróżnicowania opartych na skorygowanym indeksie Randa. Definicja tego miernika jest następująca. Niech A i B będą wynikami dwóch różnych klasyfikacji zbioru Z mającego N elementów. Przez l_A oznaczmy liczbę klas w klasyfikacji A, natomiast przez l_B – liczbę klas w klasyfikacji B; N_{ij} to liczba obiektów znajdujących się w klasie i w grupowaniu A i w klasie j w klasyfikacji B; $N_{i\bullet}$ to liczba obserwacji w klasie i w klasyfikacji A, natomiast $N_{\bullet j}$ to liczba obserwacji w klasie j w klasyfikacji B. Skorygowany indeks Randa dany jest wzorem:

$$acc(A,B) = \frac{\sum_{i=1}^{l_A}\sum_{j=1}^{l_B}\binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}, \tag{1}$$

gdzie:

$$t_1 = \sum_{i=1}^{l_A} \binom{N_{i\bullet}}{2}, \tag{2}$$

$$t_2 = \sum_{j=1}^{l_B} \binom{N_{\bullet j}}{2}, \tag{3}$$

$$t_3 = \frac{2t_1 t_2}{N(N-1)}. \tag{4}$$

Zasadniczo istnieją dwa podejścia do zagadnienia pomiaru zróżnicowania: miary dla par składowych klasyfikacji oraz dla całej klasyfikacji zagregowanej. W przypadku tej pierwszej grupy miar Hadjitodorov, Kuncheva i Todorova [2006] zaproponowali miarę uśredniającą zróżnicowanie między parami klasyfikacji składowych:

$$D_p = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \left(1 - acc(C_i, C_j)\right), \tag{5}$$

gdzie: $acc(C_i, C_j)$ – wartość skorygowanego indeksu Randa między parą klasyfikacji składowych,

M – liczba klasyfikacji składowych.

Należy zauważyć, że ze względu na to, że $acc(C_i, C_j)$ jest miarą podobieństwa, zatem $1 - acc(C_i, C_j)$ jest miarą zróżnicowania między parą klasyfikacji składowych.

W przypadku miar dla całej klasyfikacji zagregowanej, po tym, jak zostanie określona ostateczna klasyfikacja zagregowana, dla każdej klasyfikacji składowej określana jest indywidualna miara mierząca stopień różnicy między nią a klasyfikacją zagregowaną. Zatem by uzyskać ogólną miarę zróżnicowania, należy uśrednić M indywidualnych miar zróżnicowania:

$$D_{np1} = \frac{1}{M} \sum_{i=1}^{M} \left(1 - acc(C_i, C_{agr})\right). \tag{6}$$

W swoich wcześniejszych badaniach Hadjitodorov i in. [2006] odkryli, że klasyfikacja zagregowana, której elementy składowe wykazują większe rozproszenie, generalnie jest dokładniejsza niż klasyfikacja zagregowana z mniejszym rozproszeniem elementów składowych. Na podstawie tej obserwacji zaproponowali kolejną miarę dla całej klasyfikacji zagregowanej – odchylenie standardowe indywidualnych miar zróżnicowania:

$$D_{np2} = \sqrt{\frac{1}{(M-1)} \sum_{i=1}^{M} \left(1 - acc(C_i, C_{agr}) - D_{np1}\right)^2}. \tag{7}$$

Wspomniani autorzy zaproponowali jeszcze inną miarę, której konstrukcja oparta jest na następującym założeniu. Przyjmując, że klasyfikacja zagregowana jest bliska prawdziwym etykietom klas, dokładność klasyfikacji składowych może zostać określona na podstawie tego, jak bliska jest klasyfikacji ostatecznej. Zatem preferowane są wysokie wartości miary $1 - D_{np1}$. Z drugiej jednak strony zróżnicowanie klasyfikacji zagregowanej może zostać określone przez rozproszenie klasyfikacji składowych. Wysokie zróżnicowanie będzie wykazywane przez wysokie wartości miary D_{np2}. Zatem najprostsza miara będąca kompromisem dla miar $1 - D_{np1}$ oraz D_{np2} może zostać zapisana jako:

$$D_{np3} = \frac{1}{2}(1 - D_{np1} + D_{np2}) . \tag{8}$$

Ostatnią zaproponowaną miarą jest współczynnik zmienności:

$$D_{np4} = \frac{D_{np2}}{D_{np1}} . \tag{9}$$

Z eksperymentów przeprowadzonych przez wspomnianych autorów wynika jednak, że relacja między poziomem zróżnicowania mierzonym za pomocą tych miar a dokładnością ostatecznej klasyfikacji nie była zbyt jasna. Dlatego celem tego badania jest próba zastosowania miar pozycyjnych w przedstawionych miernikach zróżnicowania w miejsce zastosowanych przez autorów miar klasycznych, aby sprawdzić, czy taka modyfikacja przyniesie poprawę rezultatów. W efekcie średnia arytmetyczna we wzorach (5) i (6) zastąpiona została medianą, natomiast odchylenie standardowe we wzorze (7) zastąpiono odchyleniem ćwiartkowym.

Badanie to jest kontynuacją wcześniejszych rozważań[2], w których zastosowano miary zróżnicowania (oparte na miarach klasycznych), z zastosowaniem jednakże innych niż skorygowany indeks Randa mierników podobieństwa, a konkretnie były to indeksy: Randa, Jaccarda, Fowlkes i Mallowsa. Jednakże badania empiryczne pokazały, że zastosowanie innych indeksów nie prowadziło do wyraźnych zmian w wynikach, dlatego w badaniu tym ograniczono się do zastosowania tylko i wyłącznie skorygowanego indeksu Randa.

3. Badania empiryczne

W eksperymentach zastosowano sztucznie generowane zbiory danych, które standardowo wykorzystywane są w badaniach porównawczych w taksonomii. Wszystkie wygenerowane zostały w pakiecie `mlbench` w programie **R**. Krótka ich charakterystyka znajduje się w tab. 1.

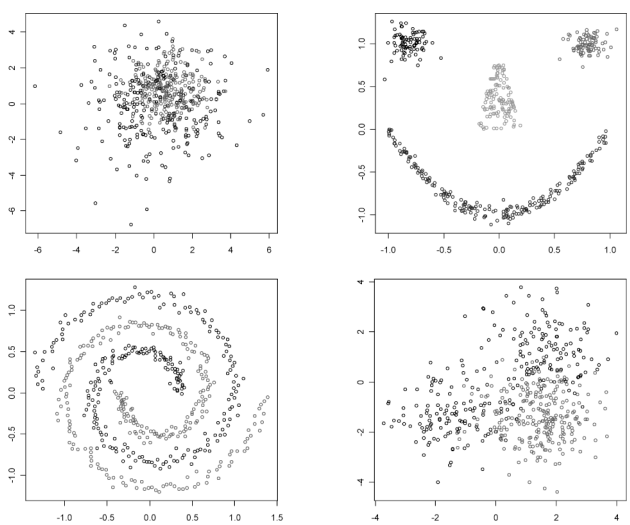
[2] Dorota Rozmus, *Analysis of diversity-accuracy relations in cluster ensemble*, w recenzji.

Tabela 1. Charakterystyka zastosowanych zbiorów danych

Zbiór danych	Liczba obiektów	Liczba cech	Liczba klas
Cuboids	500	3	4
Ringnorm	500	2	2
Smiley	500	2	4
Spirals	500	2	2
Threenorm	500	2	2

Źródło: opracowanie własne na podstawie dokumentacji programu **R**.

Zbiory *Ringnorm* oraz *Threenorm* są zbiorami dwuklasowymi wygenerowanymi w dwóch wymiarach z trudno separowalnymi klasami, dwuwymiarowe są także zbiory *Spirals* mający dwie klasy oraz *Smiley* mający cztery wyraźne klasy (rys. 1); natomiast *Cuboids* jest zbiorem czteroklasowym, wygenerowanym w trzech wymiarach (rys. 2).



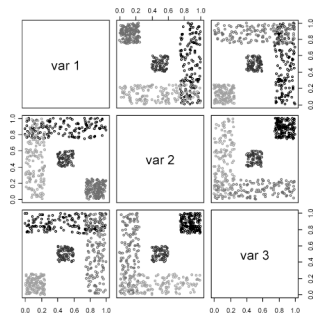
Rys. 1. Zastosowane zbiory danych; u góry – *Ringnorm* i *Smiley*, na dole – *Spirals* i *Threenorm*

Źródło: opracowanie własne.

Cały eksperyment obejmował 50 klasyfikacji zagregowanych. Każda klasyfikacja zagregowana miała 50 składowych zbudowanych za pomocą metody *k*-średnich, a wyniki agregowane były metodą *bagging* [Hornik 2005]. Całość obliczeń wykonana została w programie **R** z zastosowaniem pakietu `clue`.

Podstawowym założeniem przeprowadzonych badań jest to, że między miarami zróżnicowania i dokładności klasyfikacji zagregowanej będzie istniała na tyle jasna i oczywista relacja, że mierniki zróżnicowania będą pomocne w doborze

takich metod lub parametrów tych metod, które gwarantować będą jak najlepsze (tj. jak najdokładniejsze) wyniki klasyfikacji.



Rys. 2. Zastosowane zbiory danych – zbiór *Cuboids*

Źródło: opracowanie własne.

Przechodząc do wyników empirycznych na wstępie, należy zauważyć, że jedynym zbiorem, dla którego zaobserwowano wyraźną relację między miarami zróżnicowania a dokładnością klasyfikacji, był zbiór *Threenorm*, gdzie relacja ta w większości przypadków może zostać uznana za liniową (rys. 3). Przy czym w przypadku miar D_p i D_{np1} jest to związek o kierunku ujemnym, natomiast dla pozostałych miar – o kierunku dodatnim. Zastosowanie miar pozycyjnych w porównaniu z klasycznymi w przypadku niektórych miar pozwoliło na uściślenie relacji między zróżnicowaniem a dokładnością klasyfikacji – jest tak w przypadku miar D_p, D_{np1} i D_{np3}, w przypadku miary D_{np2} relacja ta jest znacznie mniej ścisła, natomiast w przypadku miary D_{np4} zmienia się nieznacznie kształt zależności i przypomina kształt funkcji wykładniczej. Policzone zostały także wartości współczynników korelacji liniowej, by móc porównać siłę związku między dokładnością klasyfikacji a miarami zróżnicowania z zastosowaniem miar klasycznych i pozycyjnych. Na podstawie wyników zamieszczonych w tab. 2 można zauważyć, że

Tabela 2. Wartości współczynnika korelacji liniowej Pearsona między miarami zróżnicowania a dokładnością dla zbioru Threenorm

Miara zróżnicowania	Miary klasyczne	Miary pozycyjne
D_p	–0,679	–0,690
D_{np1}	–0,869	–0,975
D_{np2}	0,933	0,389
D_{np3}	0,973	0,980
D_{np4}	0,957	0,886

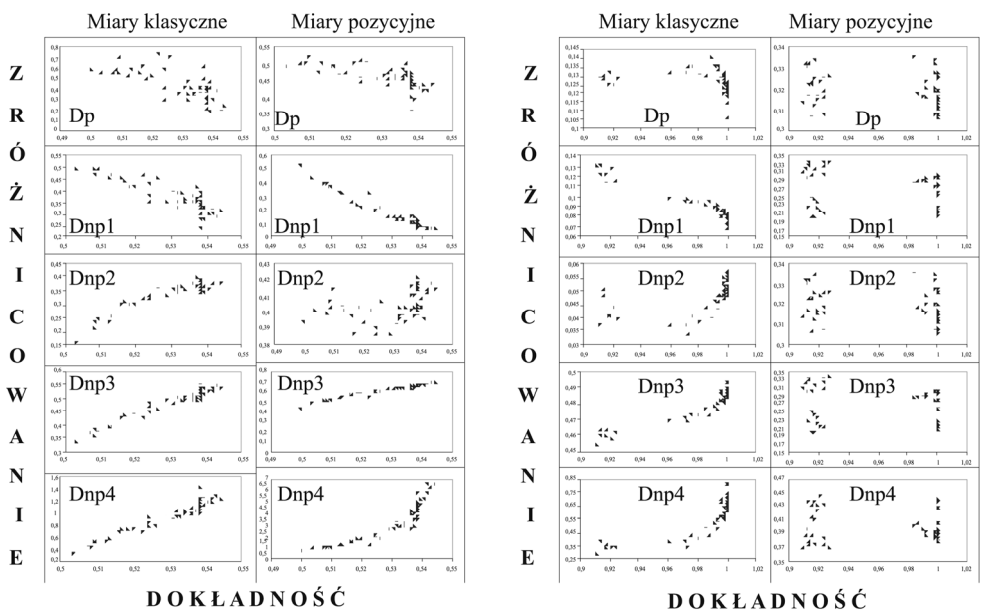
Źródło: obliczenia własne.

zastosowanie miar pozycyjnych spowodowało wzrost ścisłości związku w przypadku miar D_p, D_{np1} i D_{np3}. W przypadku miary D_{np2} widać wyraźne obniżenie stopnia zależności między miarami zróżnicowania i dokładnością klasyfikacji, co widoczne było też na rys. 3, gdzie przebieg punktów jest bardzo rozproszony i w niewielkim stopniu przypominał związek liniowy. Z oczywistych względów odnotowujemy też spadek siły związku dla miary D_{np4}.

Pozostałe zbiory danych nie dały już tak jasnej i ścisłej relacji między jakością klasyfikacji a miarami zróżnicowania, a zastosowanie miar pozycyjnych nie przyniosło poprawy rezultatów.

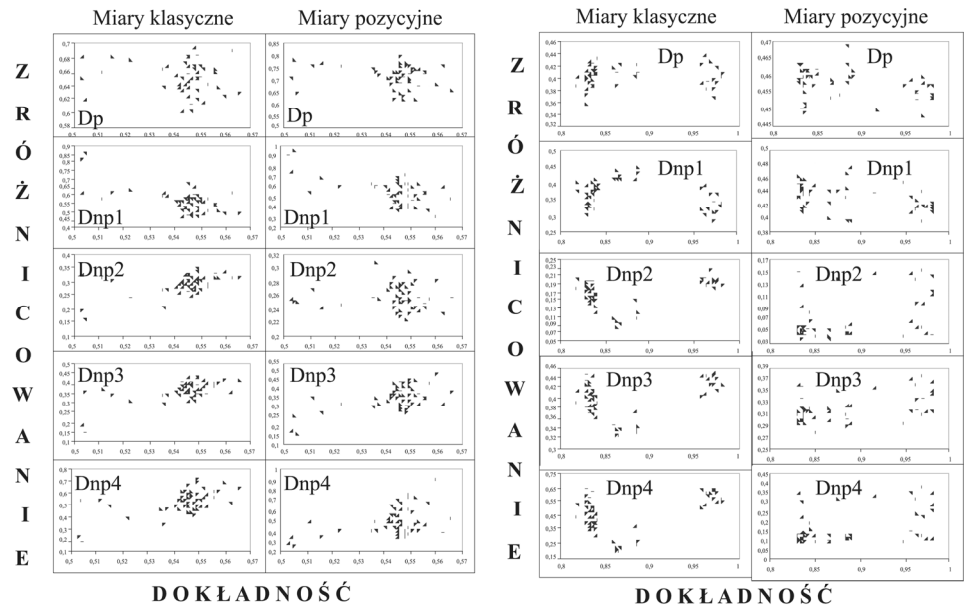
W przypadku zbioru *Cuboids* (rys. 3), o ile jeszcze miary klasyczne pozwalały zaobserwować pewną prawidłowość, którą w przypadku pierwszych dwóch miar określić by można: „mniejsze zróżnicowanie – większa dokładność", natomiast dla pozostałych miar – „większe zróżnicowanie – większa dokładność", o tyle miary pozycyjne nie wykazują już żadnej prawidłowości.

Dla zbioru *Ringnorm* w przypadku miar D_{np2}, D_{np3} oraz D_{np4}, zwłaszcza z miarami klasycznymi, widoczne są chmury punktów rozciągające się od lewego górnego rogu do prawego górnego, co mogłoby sugerować korelację dodatnią, jednakże punkty są zbyt rozproszone, by można było mówić o jakimś ścisłym związku (rys. 4).



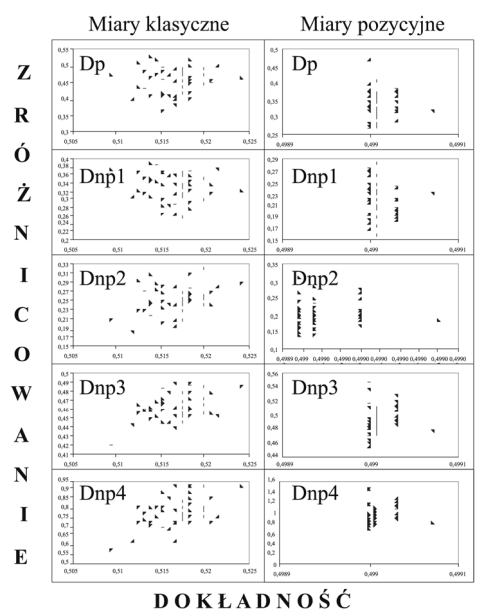
Rys. 3. Wyniki dla zbiorów *Threenorm* (po lewej) oraz *Cuboids* (po prawej)

Źródło: opracowanie własne.



Rys. 4. Wyniki dla zbiorów *Ringnorm* (po lewej) oraz *Smiley* (po prawej)

Źródło: opracowanie własne.



Rys. 5. Wyniki dla zbioru *Spirals*

Źródło: opracowanie własne.

Dla zbioru *Smiley* pierwsze dwie miary, czy to oparte na miarach klasycznych, czy na pozycyjnych, ujawniają brak jakichkolwiek relacji między zróżnicowaniem a dokładnością klasyfikacji (rys. 4). Natomiast miary D_{np2}, D_{np3} oraz D_{np4} oparte na miernikach klasycznych ujawniają, że chociaż bywa, że wysokie zróżnicowanie idzie w parze ze stosunkowo niską dokładnością, to jednocześnie widać także, że wysokie zróżnicowanie jest niezbędne dla wysokiej dokładności klasyfikacji. Miary pozycyjne natomiast zaburzają ten schemat.

Bywało i tak, jak w przypadku zbioru *Spirals*, że zastosowane miary, czy to oparte na miernikach klasycznych, czy pozycyjnych, nie wykazywały żadnej relacji między zróżnicowaniem składowych a dokładnością klasyfikacji zagregowanej (rys. 5).

4. Wnioski

Podsumowując to badanie, należy zauważyć, że powszechnie uznaje się, że klasyfikacje zagregowane charakteryzujące się większym zróżnicowaniem elementów składowych dają większą poprawę dokładności klasyfikacji w porównaniu z mniej zróżnicowanymi; jednocześnie z zagadnienia dyskryminacji wynika także, że relacja między zróżnicowaniem i dokładnością nie jest jasna i bezpośrednia [Kuncheva, Whitaker 2003]. Jako że zróżnicowanie nie jest ściśle zdefiniowanym terminem, istnieje wiele możliwości pomiaru. W badaniu wykorzystano koncepcję pięciu miar zaproponowanych w literaturze, z pewną jednakże modyfikacją polegającą na zastąpieniu mierników klasycznych miernikami pozycyjnymi. Był to kolejny pomysł na drodze próby znalezienia miar pozwalających określić relację między zróżnicowaniem a dokładnością klasyfikacji w podejściu zagregowanym w taksonomii. Na podstawie wyników empirycznych jednakże okazuje się, że modyfikacja ta nie prowadzi do uzyskania lepszych efektów. Wyjątek stanowi tylko zbiór *Threenorm* dla miar D_p, D_{np1} i D_{np3}.

Literatura

Breiman L., *Bagging predictors*, „Machine Learning" 1996, 26(2).

Fern X.Z., Brodley C.E., *Random projection for high dimensional data clustering: a cluster ensemble approach*, „Proceedings of the 20th International Conference on Machine Learning", ICML, Washington, DC 2003.

Fred A., *Finding Consistent Clusters in Data Partitions*, [w:] *Proceedings of the International Workshop on Multiple Classifier Systems*, F. Roli, J. Kittler (red.), LNCS 2364, 2002.

Fred A., Jain A.K., *Data clustering using evidence accumulation*, „Proceedings of the 16th International Conference on Pattern Recognition", ICPR, Canada 2002, 276-280.

Freund Y., *Boosting a weak learning algorithm by majority*, „Proceedings of the 3rd Annual Workshop on Computational Learning Theory" 1990.

Hadjitodorov S.T., Kuncheva L.I., Todorova L.P., *Moderate diversity for better cluster ensembles*, „Information Fusion" 2006, 7.

Hornik K., *A CLUE for CLUster Ensembles*, „Journal of Statistical Software" 2005, 14.

Kuncheva L.I., *That elusive diversity in classifier ensembles*, „Lecture Notes in Computer Science", Springer-Verlag, Mallorca, Spain 2003, vol. 2652.
Kuncheva L.I., Hadjitodorov S.T., *Using diversity in cluster ensembles*, „Proceedings of IEEE International Confference on Systems, Man and Cybernetics" 2004.
Kuncheva L.I., Whitaker C.J., *Measures of diversity in classifier ensembles*, „Machine Learning" 2003, 51.
Strehl A., Ghosh J., *Cluster ensembles – a knowledge reuse framework for combining partitionings*, „Journal of Machine Learning Research" 2002, 3.
Tsymbal A., Pechenizkiy M., Cunningham P., *Diversity in ensemble feature selection*, „Technical Report", Trinity College, Dublin 2003.

APPLYING OF POSITION MEASURES FOR STUDY OF RELATIONSHIP BETWEEN DIVERSITY AND ACCURACY IN CLUSTER ENSEMBLE

Summary: Ensemble approach has been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. Recently, analogous techniques for cluster analysis have been suggested.
Diversity within an ensemble is of vital importance for its success. The main aim of this research is a trial to apply position measures in proposed in literature diversity measures [Hadjitodorov et al. 2006] in order to check the relationship between diversity and accuracy in cluster ensemble and compare the results with as yet applied ways of measuring it.