PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Dorota Rozmus
Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE STABILNOŚCI ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA MACIERZY WSPÓŁWYSTĄPIEŃ

Streszczenie: Podejście zagregowane (wielomodelowe) dotychczas z dużym powodzeniem stosowane było w dyskryminacji w celu podniesienia dokładności klasyfikacji. W ostatnich latach analogiczne propozycje pojawiły się w taksonomii, aby zapewnić większą poprawność i stabilność wyników grupowania. Stabilność algorytmu taksonomicznego w odniesieniu do niewielkich zmian w zbiorze danych czy też parametrów algorytmu jest pożądaną cechą algorytmu. Głównym celem artykułu jest porównanie stabilności zagregowanych algorytmów taksonomicznych opartych na macierzy współwystąpień oraz zbadanie relacji, jakie zachodzą między stabilnością a dokładnością.

Słowa kluczowe: taksonomia, podejście zagregowane, stabilność, dokładność klasyfikacji, macierz współwystąpień.

1. Wstęp

Pierwotnie podejście zagregowane (wielomodelowe) z dużym powodzeniem stosowane było w dyskryminacji i regresji w celu podniesienia dokładności predykcji. Zasadnicza idea tego podejścia polega na tym, że w pierwszym kroku budowane są liczne, różniące się między sobą, pojedyncze modele, które następnie za pomocą różnych operatorów łączy się w model zagregowany. W klasyfikacji najczęściej stosowanym operatorem jest głosowanie majoryzacyjne, co oznacza, że wybiera się tę klasę, która najczęściej wskazywana była przez pojedyncze modele; natomiast w regresji najczęściej stosuje się uśrednianie wartości teoretycznych zmiennej y. Wśród najbardziej znanych metod agregacji wymienić należy: *bagging* [Breiman 1996], który oparty jest na losowaniu kolejnych prób boostrapowych, oraz *boosting* [Freund 1990] polegający na nadawaniu wyższych wartości wag błędnie sklasyfikowanym obiektom.

W ostatnich latach analogiczne propozycje pojawiły się także w taksonomii, aby zapewnić większą poprawność i stabilność wyników klasyfikacji [Fern, Brodley 2003, Fred 2002, Fred, Jain 2002, Strehl, Ghosh 2002]. Zagadnienie agregacji w taksonomii może zostać sformułowane następująco: mając wyniki wielokrotnie przeprowadzonego grupowania, znajdź zagregowany podział ostateczny o lepszej jakości. Liczne badania w tej dziedzinie ustanowiły już nowy obszar w tradycyjnej taksonomii. Istnieją liczne możliwości zastosowania idei podejścia zagregowanego w dziedzinie uczenia bez nauczyciela, wśród których jako najpopularniejsze należy wymienić:

1) łączenie wyników grupowania uzyskanych za pomocą różnych metod,

2) uzyskanie różniących się między sobą klasyfikacji z zastosowaniem różnych podzbiorów danych, np. poprzez losowanie bootstrapowe,

3) stosowanie różnych podzbiorów zmiennych,

4) wielokrotne zastosowanie tego samego algorytmu z różnymi wartościami parametrów lub punktami startowymi (np. losowo wybranymi zalążkami skupień w metodzie *k*-średnich).

Pożądaną cechą algorytmu taksonomicznego jest, by wykazywał on stabilność, a więc był odporny na niewielkie zmiany w zbiorze danych, czy też wartości parametrów tego algorytmu. Wiadomo jednakże również, że kluczem do sukcesu podejścia zagregowanego jest zróżnicowanie klasyfikacji składowych. Klasyfikacja zagregowana, która zbudowana została na różniących się między sobą elementach składowych, jest bardziej dokładna i stabilna niż pojedyncze metody taksonomiczne.

W niniejszym badaniu uwagę skupiono na stabilności metod taksonomicznych. Głównym celem tego artykułu jest porównanie stabilności zagregowanych algorytmów taksonomicznych, a także relacji między stabilnością i dokładnością; przy czym wzięta zostanie pod uwagę tylko specyficzna klasa metod agregacji, która oparta jest na tzw. macierzy współwystąpień.

2. Metoda agregacji oparta na macierzy współwystąpień

Generalnie rzecz ujmując, w taksonomii istnieją trzy źródła tego sposobu agregacji. Pierwszym jest podejście oparte na opisie zbioru obiektów poprzez podobieństwo (bądź niepodobieństwo), które zaproponowane zostało w dyskryminacji. W podejściu tradycyjnym modele dyskryminacyjne budowane są na podstawie zbioru danych, który zawiera zmienne charakteryzujące poszczególne obiekty. Alternatywą dla tego klasycznego opisu obserwacji może być podejście oparte na macierzy podobieństwa (odległości) między obiektami, które zaproponowane zostało przez Pekalską i Duin [2000]. W metodzie tej obiekty opisywane są przez pewną miarę obrazującą stopień podobieństwa (bądź niepodobieństwa) między obserwacjami ze zbioru danych. Model zatem jest budowany na macierzy podobieństwa bądź odległości, którą traktuje się jako zbiór danych opisujących poszczególne obiekty.

Drugie źródło to zaproponowana przez Fred i Jain [2002] idea łączenia wyników wielokrotnie dokonanego grupowania w celu konstrukcji tzw. macierzy współwystąpień. W podejściu tym pod uwagę bierze się jednoczesne wystąpienie pary obiektów w tej samej grupie, które traktuje się jako wskazówkę istnienia związku między nimi. Pierwotny zbiór obserwacji zatem przekształcany jest w $n \times n$-wymiarową macierz, która opisuje podobieństwo między obiektami. Ostatecznego grupowania dokonuje się na podstawie uzyskanej macierzy współwystąpień, która traktowana jest jako macierz odległości między obiektami.

I trzecie źródło to obiecujące rezultaty badań taksonomicznych [Kuncheva i in. 2006], w których macierz współwystąpień potraktowana została jako macierz danych.

W badaniu tym zastosowano zatem dwuetapowe podejście. Najpierw, po uzyskaniu klasyfikacji składowych, skonstruowana była macierz współwystąpień (czyli jest to etap agregacji), służąca potem jako macierz danych dla różnych metod klasyfikacji, których stabilność była badana. Dokładniej, krok pierwszy, służący konstrukcji macierzy współwystąpień, może zostać sformułowany następująco:

Wielokrotna klasyfikacja. Dla założonej liczby składowych *C*, wchodzących w skład macierzy współwystąpień, dokonaj grupowania obiektów, np. za pomocą metody *k*-średnich, uzyskując różniące się między sobą rezultaty dzięki losowo wybranym zalążkom skupień.

Agregacja. Podstawą tego podejścia jest założenie, że obiekty należące do tej samej grupy najprawdopodobniej będą znajdowały się w tej samej klasie wśród tych *C* podziałów. Traktując zatem współwystąpienie pary obiektów w tej samej grupie jako wskazówkę istnienia związku między nimi, wyniki klasyfikacji uzyskane dzięki wielokrotnie zastosowanej metodzie *k*-średnich są przekształcane w $n \times n$-wymiarową macierz współwystąpień zgodnie ze wzorem:

$$co\_assoc(a,b) = votes_{ab} \tag{1}$$

gdzie $votes_{ab}$ zlicza, ile razy para obiektów *a* i *b* zaliczona została do tej samej grupy, wśród tych *C* składowych klasyfikacji.

Ostateczny podział. W celu uzyskania ostatecznego podziału, zastosuj dowolny algorytm taksonomiczny do skonstruowanej wcześniej macierzy współwystąpień, traktując ją jak macierz danych.

3. Miary stabilności i dokładności

W celu zbadania stabilności i dokładności zastosowano koncepcję miar zaproponowanych przez Kunchevą i Vetrova [2006]. Wszystkie te mierniki oparte są na skorygowanym indeksie Randa (*AR*).

1. Stabilność dla par klasyfikacji zagregowanych (*pairwise ensemble stability*):

$$S_{agr} = \frac{2}{K \cdot (K-1)} \sum_{\substack{1 \le k,l \le K \\ k<l}}^{K} AR(P_k^{agr}, P_l^{agr}) \tag{2}$$

gdzie: K – liczba klasyfikacji zagregowanych,
AR – skorygowany indeks Randa,
P_k^{agr} – klasyfikacja na podstawie *k*-tej klasyfikacji zagregowanej,
P_l^{agr} – klasyfikacja na podstawie *l*-tej klasyfikacji zagregowanej.

Miara ta ocenia stabilność klasyfikacji zagregowanych poprzez ocenę podobieństwa wyników grupowania, które na ich podstawie zostały uzyskane.

2. Przeciętna dokładność klasyfikacji zagregowanej (*average ensemble accuracy*):

$$A_{agr} = \frac{1}{K} \sum_{k=1}^{K} AR(P_k^{agr}, P^T), \tag{3}$$

gdzie P^T to rzeczywiste etykiety klas.

Miara ta jest uśrednioną po wszystkich klasyfikacjach zagregowanych miarą dokładności i mierzy podobieństwo między ostateczną klasyfikacją zagregowaną a prawdziwymi etykietami klas.

4. Badania empiryczne

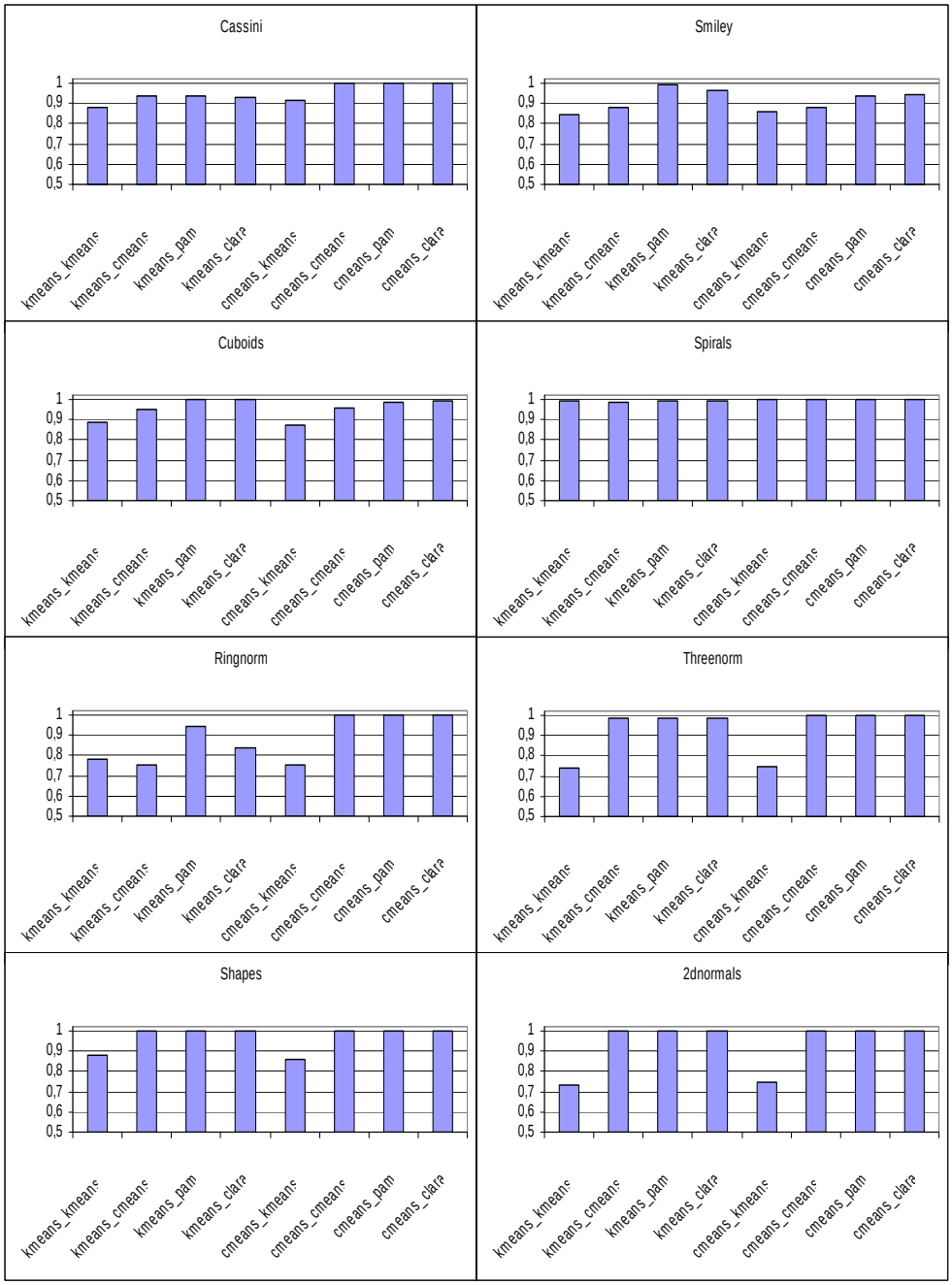
W badaniach zastosowano sztucznie generowane zbiory danych, które standardowo wykorzystywane są w badaniach porównawczych w taksonomii[1]. Są to takie zbiory, w których przynależność obiektów do klas jest znana. Ich krótka charakterystyka znajduje się w tab. 1.

Tabela 1. Charakterystyka zastosowanych zbiorów danych

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

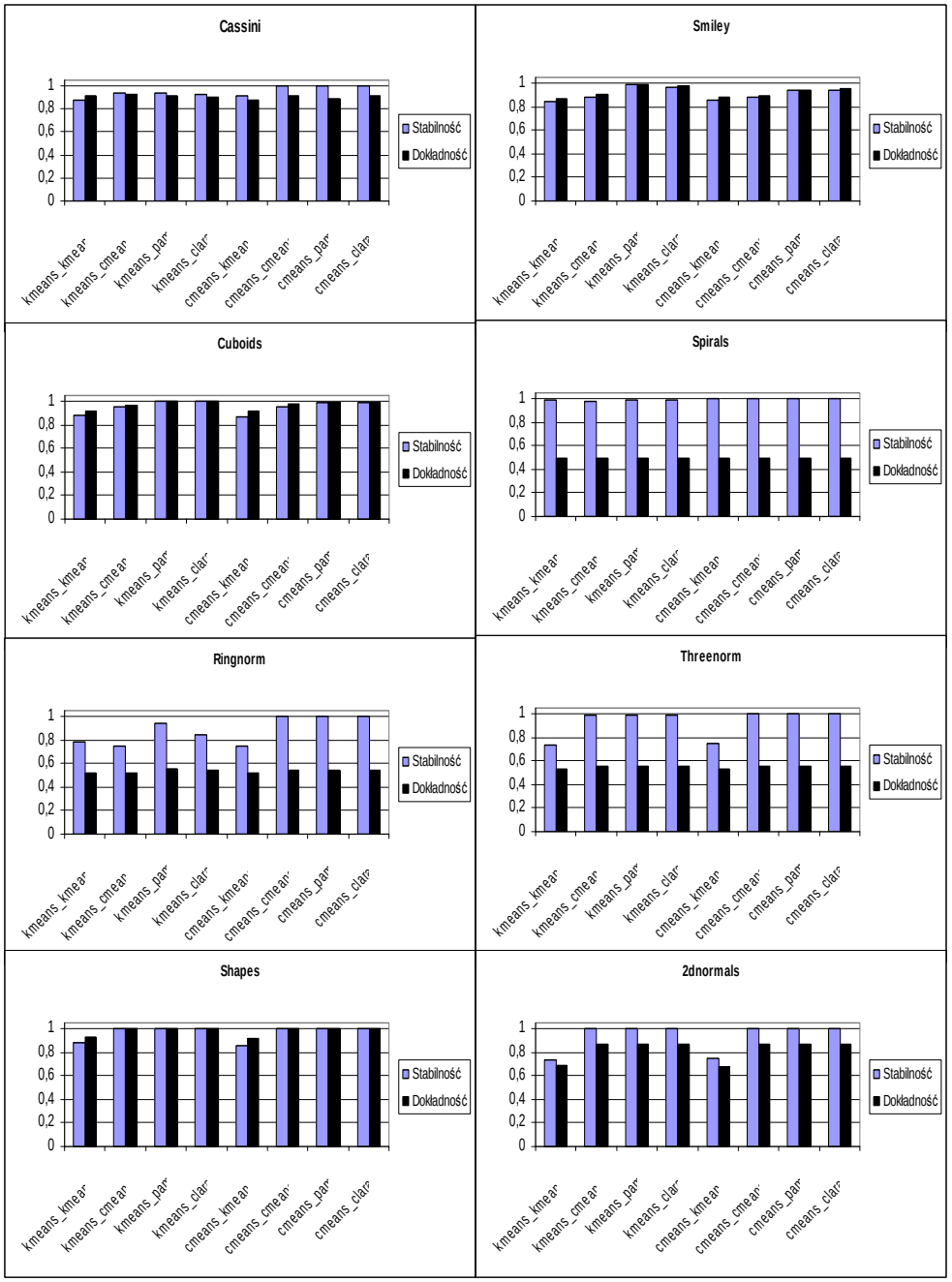
Źródło: opracowanie własne.

[1] Zaczerpnięte zostały z pakietu `mlbench` z programu **R**.



Rys. 1. Stabilność poszczególnych metod dla różnych zbiorów danych

Źródło: opracowanie własne.



Rys. 2. Relacje między stabilnością a dokładnością poszczególnych metod dla różnych zbiorów danych

Źródło: opracowanie własne.

Do konstrukcji macierzy współwystąpień zastosowano metodę *k*-średnich oraz metodę *c*-średnich, która jest rozmytą wersją metody *k*-średnich, opracowaną przez Bezdeka [1981]. W procesie konstrukcji macierzy współwystąpień liczbę składowych klasyfikacji przyjęto równą 10, natomiast parametry *k* i *c* były równe liczbie klas.

Wśród zastosowanych później metod podziału skonstruowanych macierzy współwystąpień zastosowano metodę: *k*-średnich, *c*-średnich, *k*-medoidów [Kaufman, Rousseeuw 1990] oraz clara [Kaufman, Rousseeuw 1990].

Dla każdego zbioru danych określano 10 klasyfikacji zagregowanych, których stabilność oraz dokładność w następnym kroku była mierzona za pomocą miar (2) i (3).

Ponadto dla każdego zbioru danych badania powtórzono 20 razy w celu uzyskania bardziej wiarygodnych i dokładnych rezultatów. Wszystkie obliczenia wykonane zostały w programie **R**.

Dla wszystkich zbiorów (rys. 1) najbardziej stabilna okazała się metoda cmeans_cmeans, cmeans_pam oraz cmeans_clara (z wyjątkiem zbiorów *Cuboids* oraz *Smiley*, gdzie najskuteczniejsza była metoda kmeans_pam)[2]. Stosunkowo słabą stabilnością charakteryzuje się metoda cmeans_kmeans, z wyjątkiem zbioru *Spirals*. Wśród metod, gdzie macierz współwystąpień budowana była za pomocą metody *k*-means, najmniej stabilna w przypadku wszystkich zbiorów danych okazała się metoda kmeans_kmeans (z wyjątkiem zbiorów *Spirals* i *Ringnorm*, gdzie jeszcze słabsza była metoda kmeans_cmeans). Dla większości zbiorów stosunkowo stabilne okazały się także metody kmeans_pam i kmeans_clara.

Wykresy na rys. 2, pokazujące relacje między miarami stabilności i dokładności, pozwalają stwierdzić, że trudno określić generalną zależność czy relację. Dla zbiorów *Shapes*, *Smiley* oraz *Cuboids* miary stabilności i dokładności przybierają bardzo przybliżone wartości (z wyjątkiem metod kmeans_kmeans i cmeans_kme ans dla zbioru *Shapes* oraz dodatkowo kmeans_cmeans i cmeans_cmeans dla zbiorów *Cuboids* i *Smiley*). Dla zbiorów *2dnormals* oraz *Cassini* dokładność wszystkich metod kształtuje się nieco poniżej poziomu stabilności, z wyjątkiem metody kmeans_kmeans dla zbioru *Cassini*. Natomiast dla pozostałych zbiorów dokładność wszystkich metod kształtuje się na niemalże takim samym poziomie, natomiast stabilność − zależnie od metody, np. dla zbioru *Spirals* różnice są niewielkie, a dla pozostałych już bardziej znaczące[3].

5. Zakończenie

Przechodząc do sformułowania wniosków ostatecznych, należy zauważyć, że wiele algorytmów taksonomicznych, w tym także tych, w których wykorzystuje się podejście

[2] Pierwszy element nazwy odnosi się do metody konstrukcji macierzy współwystąpień, a drugi do sposobu jej ostatecznego podziału.

[3] Ze względu na to, że punktem zainteresowania badania była relacja między miarami stabilności i dokładności, zdecydowano się także na analizę tych zależności w przypadku, gdy dokładność klasyfikacji nie była zbyt wysoka.

zagregowane, zawiera element losowości. W związku z tym pożądaną cechą algorytmu taksonomicznego jest stabilność uzyskiwanych wyników w odniesieniu do niewielkich zmian w wartościach parametrów tych algorytmów czy też w zbiorze danych.

Wiadomo również, że zróżnicowanie klasyfikacji składowych w podejściu zagregowanym przyczynia się do sukcesu tego podejścia. Oczekuje się, że grupowanie zagregowane, budowane na różniących się między sobą składowych, będzie dawało bardziej stabilne rezultaty niż pojedyncze klasyfikacje. Celem niniejszego badania było porównanie stabilności zagregowanych algorytmów taksonomicznych opartych na macierzy współwystąpień, a także relacji między stabilnością i dokładnością. Z przeprowadzonych badań wynika, że wśród wszystkich metod najczęściej najbardziej stabilna okazywały się metody cmeans_cmeans, cmeans_pam oraz cmeans_clara oraz że nie można określić jednoznacznej relacji między stabilnością i dokładnością. Dla niektórych zbiorów danych stabilność i dokładność kształtują się na zbliżonym poziomie, a dla niektórych stwierdza się brak jakiegokolwiek związku między nimi.

Literatura

Bezdek J.C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York 1981.

Breiman L., *Bagging predictors*, „Machine Learning" 1996, 26(2), s. 123-140.

Fern X.Z., Brodley C.E., *Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach*, Proceedings of the 20th International Conference of Machine Learning, Washington 2003, s. 186-193.

Fred A., *Finding Consistent Clusters In Data Partitions*, [w:] F. Roli, J. Kittler (red.), *Proceedings of the International Workshop on Multiple Classifier Systems*, Cagliari 2002, 309-318.

Fred A., Jain A.K., *Data Clustering Using Evidence Accumulation*, Proceedings of the 16th International Conference on Pattern Recognition , ICPR, Canada, 2002, s. 276-280.

Freund Y., *Boosting a Weak Learning Algorithm by Majority*, Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Rochester 1990, s. 202-216.

Kaufman L., Rousseeuw P.J., *Finding Groups in Data*: *An Introduction to Cluster Analysis*, Wiley, New York 1990.

Kuncheva L.I., Hadjitodorov S.T., Todorova L.P., *Experimental comparison of cluster ensemble methods*, Proceedings of the 19th International Conference on Information Fusion, Florence 2006, s. 1-7.

Kuncheva L., Vetrov D., *Evaluation of stability of k-means cluster ensembles with respect to random initialization*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" 2006, vol. 28, no. 11, s. 1798-1808.

Pekalska E., Duin R.P.W., *Classifiers for dissimilarity-based pattern recognition*, A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquezar, A.K. Jain, J. Kittler J. (red.), Proceedings of the 15th International Conference on Pattern Recognition, IEEE Computer Society, Press, Los Alamitos 2000, s. 12-16.

Strehl A., Ghosh J., *Cluster ensembles − A knowledge reuse framework for combining multiple partitions*, „Journal of Machine Learning Research" 2002, 3, s. 583-618.

COMPARISON OF STABILITY OF CLUSTER ENSEMBLES BASED ON CO-OCCURRENCE DATA

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 in order to increase classification accuracy, robustness and stability of the clustering solutions. The stability of a clustering algorithm with respect to small perturbations of data or parameters of the algorithm is a desirable quality of the algorithm. In the article we look at the stability of the ensemble methods. This paper carries out an experimental study to compare the stability of cluster ensembles based on co-occurrence matrix and it also looks at the relationship between stability and accuracy measures.